UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Programação em Excel para Estatística: Modelo Linear e Extensões

Catarina Nunes Valente

Mestrado em Matemática Aplicada à Economia e Gestão

Trabalho de Projeto orientado por:
Prof.ª Doutora Teresa Alpuim

2017

AGRADECIMENTOS

Este projeto foi desenvolvido sob a orientação da Professora Teresa Alpuim, no âmbito do Mestrado em Matemática Aplicada à Economia e Gestão, a quem agradeço toda a disponibilidade prestada desde o início do meu percurso académico nesta instituição.

Deixo um agradecimento especial à minha família, em particular aos meus pais, pelo seu apoio incondicional nos bons e nos maus momentos, por me terem dado acesso a uma educação de qualidade, e por me incentivarem a construir o meu futuro.

Agradeço ainda a todas as pessoas que de alguma forma contribuíram para a realização deste projeto com as suas palavras de motivação e críticas construtivas.

RESUMO

No nosso dia-a-dia podem ser encontrados vários exemplos de relações entre várias variáveis, sejam elas físicas, sociais, económicas, políticas ou de outro tipo. Por exemplo, existe uma clara e direta relação entre a altura de uma pessoa e o seu peso. Outros exemplos de relações entre variáveis podem ser as que existem entre o rendimento e o consumo de uma pessoa, ou um país, ou a propensão para o incumprimento de uma linha de crédito e a situação financeira de uma empresa.

Uma forma simples e relevante de medir a relação entre variáveis é o conceito de análise de regressão. A análise de regressão pretende identificar a relação entre uma variável dependente e uma ou mais variáveis independentes.

Hoje em dia já existem ferramentas que permitem a um analista, de forma sofisticada, identificar relações entre variáveis e desenvolver modelos de regressão preditivos e robustos. Contudo, para além de necessitarem de licenças para a sua utilização, estas ferramentas nem sempre estão disponíveis ao utilizador comum e são dispendiosas.

Atualmente, qualquer pessoa tem facilmente acesso ao *software* Microsoft Office Excel, ou a uma ferramenta semelhante, pelo que, desta forma, as ferramentas desenvolvidas em Excel são acessíveis a todos. Por outro lado, a possibilidade de utilização das funções e métodos do Excel conjuntamente com procedimentos já programados aumenta significativamente as capacidades de análise específicas para cada caso sem perda de eficiência, por via das metodologias acessíveis num pacote.

Assim, pretende-se desenvolver ferramentas interativas e fáceis de utilizar capazes de auxiliar na identificação de relações entre variáveis, e que permitem efetuar ajustamentos de modelos estatísticos, em particular, modelos de regressão linear e modelos de regressão logística, em *Visual Basic for Applications* (VBA) do Microsoft Office Excel. Estes processos poderão então vir a ser utilizados quer por estudantes que estejam a aprender a desenvolver modelos num contexto académico, como por um profissional analista que pretenda desenvolver modelos num contexto real.

Palavras-chave: Modelos Estatísticos, Excel, VBA, Regressão Linear Múltipla, Regressão Logística.

ABSTRACT

In everyday life, we can find plenty of examples of relationships between variables, weather physical, social, economic, political or otherwise. For example, there is a clear and direct relationship between the height of a person's body and their weight. Other examples of relationships between variables could be the ones between someone's income and their expenses, or the tendency to have credit lines defaulting and the financial situation of a company.

One simple and meaningful way to measure the relationships among variables is the statistical concepts of regression analysis. Regression analysis involves identifying the relationship between a dependent variable and one or more independent variables.

Nowadays there are tools that allow for an analyst, in a sophisticated manner, to identify relationships between variables and develop robust predictive regression models. However, besides the need for a user license, these kinds of tools are not accessible to the common user and are very expensive.

Currently, anyone can have easy access to the Microsoft Office Excel Software, or a similar tool, and so, any program developed in Excel is accessible to everyone. On the other hand, the ability to use Excel functions in addition to the procedures programmed greatly enhances the specific analysis capabilities for each individual scenario without losing efficiency, using the methodologies developed in the package.

This way, this project has the objective to develop interactive and intuitive tools that can help identifying relationships between variables, and allow the development of statistical models, specifically, linear regression models and logistic regression models, in Visual Basic for Applications (VBA) of Microsoft Office Excel. Both students learning to develop models in an academic context, and professional analysts developing models in a real environment could then use these processes.

Key words: Statistical Models, Excel, VBA, Multiple Linear Regression, Logistic Regression.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

INTRODUÇÃO

Hoje em dia, a regressão linear e não-linear são das ferramentas estatísticas mais utilizadas na prática e em quase todas as áreas do conhecimento científico e técnico. Com efeito, seja porque se pretende saber a forma como uma determinada variável influencia outra ou porque se pretende obter previsões de uma quantidade em função de outras mais facilmente mensuráveis, a regressão linear múltipla e as suas extensões ocupam um lugar de destaque entre todas as metodologias estatísticas com maior utilização na prática. Por outro lado, as folhas de cálculo têm também uma utilização generalizada e constituem uma forma acessível, mesmo para os utilizadores com poucos conhecimentos de programação, de organizar e analisar dados, otimizar procedimentos, estimar valores importantes e executar uma grande diversidade de análises e procedimentos particularmente úteis na gestão de empresas e organizações.

O objetivo deste trabalho é tornar acessível ao utilizador de folhas de cálculo, um conjunto de procedimentos estatísticos de grande utilidade na gestão empresarial, em particular, na gestão de instituições financeiras.

Em instituições financeiras os modelos preditivos são um ponto comum entre vários departamentos de uma instituição e podem estar presentes em áreas como a gestão de clientes, marketing e risco (risco de crédito e risco de mercado), para nomear alguns. Em particular na área de risco de crédito, os modelos de regressão são muito usados no momento da concessão de um crédito com o objetivo de tentar prever se o cliente será complacente com as suas responsabilidades financeiras num futuro próximo. Inclusivamente, quando os resultados do modelo são favoráveis à concessão do crédito, estes podem ser usados para definir algumas condições da operação, por exemplo, taxas de juros e *spreads*.

Atualmente, o EXCEL providencia um conjunto de funções pré-definidas muito úteis para a análise estatística bem como diversas macros que permitem efetuar os procedimentos de estatística mais simples. Este projeto pretende adicionar a essas ferramentas já existentes, um conjunto de programas em *Visual Basic for Applications* (VBA) que permitem construir, ajustar e analisar modelos de regressão linear e regressão logística, permitindo também, numa primeira fase, analisar as potenciais variáveis a incluir nos modelos de uma forma mais ou menos automatizada. Ao longo do projeto, procurou-se encontrar um compromisso entre a automatização dos processos e a sua flexibilidade, permitindo ao utilizador fazer as escolhas sobre o método de construção do modelo que lhe parecerem mais adequadas ao caso específico em estudo. As folhas de cálculo apresentam grandes vantagens no que respeita à organização de dados, além disso, permitem análises mais finas e detalhadas num ambiente interativo. O objetivo é juntar esse tipo de funcionalidades com procedimentos estatísticos mais avançados.

Pressupõe-se que o utilizador destes programas esteja minimamente familiarizado com alguns conceitos de estatística, nomeadamente, testes de hipóteses, intervalos de confiança, níveis de significância, modelos de regressão linear e regressão logística e análise matricial. Adicionalmente, para uma análise mais detalhada dos procedimentos, é recomendado que o utilizador tenha alguns conhecimentos de programação, e em particular que esteja familiarizado com a ferramenta do Microsoft Excel VBA, uma vez que foi nesta plataforma que os procedimentos foram desenvolvidos.

A estrutura do relatório inclui uma primeira parte onde se apresentam com algum rigor os conceitos básicos para os quais, em seguida, se faz o acompanhamento da implementação em VBA, procurando assim dar o suporte teórico ao potencial utilizador dos programas informáticos desenvolvidos. É ainda um objetivo que este documento possa servir de apoio a quem se esteja a iniciar no desenvolvimento de modelos estatísticos e espera-se, também, que este documento possa servir de guia de utilização às ferramentas desenvolvidas.

Note-se que o enquadramento teórico deste documento se encontra muito em linha com os apontamentos disponibilizados pela Prof.ª Teresa Alpuim para a cadeira de Modelos Lineares do Mestrado em Matemática Aplicada à Economia e Gestão da Faculdade de Ciências da Universidade de Lisboa, no ano letivo de 2016/2017.

Finalmente, é relevante notar que os procedimentos aqui apresentados foram implementados e testados em Sistema Operativo Windows e Microsoft Office 2013 e Microsoft Office 2010, e que não foi possível apresentar as versões dos programas para outros Sistemas Operativos.

1ª PARTE – ENQUADRAMENTO TEÓRICO

Quando se inicia o desenvolvimento de um modelo existem diversas variáveis que podem surgir como sendo explicativas. Contudo, também é fácil obter variáveis que, apesar de teoricamente fazerem sentido que estejam no modelo, estatisticamente podem não ter relevância suficiente, ou podem existir variáveis que estejam relacionadas entre si. Este é um tema que tem que ser abordado com a devida cautela uma vez que a multicolinearidade tende a enviesar os resultados dos algoritmos que irão ser desenvolvidos neste trabalho. Adicionalmente, a existência de um grande número de variáveis a serem consideradas no modelo, torna o processo de desenvolvimento demasiado moroso.

Desta forma, após terem sido obtidas todas as variáveis que o utilizador pretende utilizar no desenvolvimento do seu modelo, devem ser analisadas as suas propriedades estatísticas. De seguida serão analisados alguns exemplos de análises preliminares ao conjunto de potenciais variáveis explicativas que permitirão obter um modelo mais robusto e eficaz.

1. MODELOS DE REGRESSÃO LINEAR

A análise de regressão linear tem como principal objetivo encontrar as equações lineares que relacionam uma variável que se pretende estudar com outras variáveis mais fáceis de medir ou de obter, com o intuito de fazer previsões e inferir sobre valores futuros. Para isso é necessário identificar a existência de relações entre a variável em estudo, dita variável dependente, e as outras variáveis, as quais se chamam variáveis independentes.

Numa primeira fase, deve ser definida uma hipótese para explicar a relação pretendida, e são calculados os estimadores dos parâmetros do modelo de regressão. Posteriormente, devem ser aplicados alguns testes estatísticos para determinar se o modelo é satisfatório. Se o modelo permitir descrever bem a amostra, e tendo disponíveis novos valores para as variáveis independentes, os estimadores dos parâmetros do modelo da regressão podem ser usados para prever valores futuros da variável dependente.

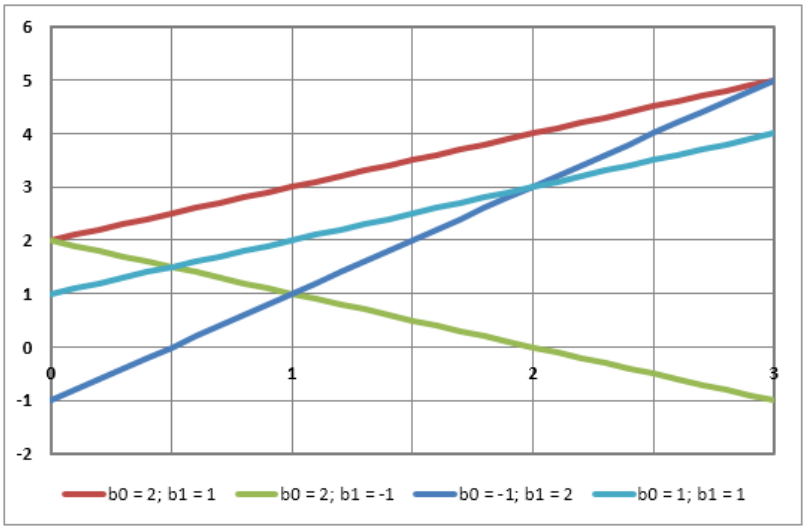

Figura 1.1 – Função Linear: $\boldsymbol{y = b_0 + b_1 x_1}$

Sem perda de generalidade, pode dizer-se que uma variável aleatória y segue um modelo de regressão linear se puder ser escrita como uma combinação linear de outras variáveis, que são tratadas como constantes, e um erro aleatório, ou seja:

$$\mathrm{y} = b_0 + b_1\mathrm{x}_1 + \cdots + b_k\mathrm{x}_k + \varepsilon, \tag{1.1}$$

em que $b_j, j = 0, \dots, k$, são os coeficientes que, em geral, se pretende estimar a partir de um conjunto de observações da variável dependente y_i e das variáveis independentes $x_{i,j}$. Assim, um conjunto de n observações do modelo de regressão linear múltipla pode ser escrito como:

$$y_i = \sum_{j=0}^{k} x_{i,j} b_j + \varepsilon_i,\ i = 1, \dots, n \tag{1.2}$$

em que os termos de erro, ε_i, são variáveis aleatórias que verificam as condições de *Gauss-Markov*, nomeadamente:

1. $\mathbb{E}(\varepsilon_i) = 0$;
2. $var(\varepsilon_i) = \sigma^2$;
3. $\mathbb{E}(\varepsilon_i \varepsilon_w) = 0, \text{se } i \neq w$,

com $i = 1, \dots n$, $w = 1, \dots n$ e $j = 0, \dots, k$.

O modelo de regressão linear múltipla também pode ser apresentado utilizando notação matricial, isto é, considerando $\mathbf{Y} = \mathbf{Xb} + \boldsymbol{\varepsilon}$, onde:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{1,0} & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,0} & x_{2,1} & x_{2,1} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n,0} & x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_0 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} \text{ e } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \text{onde } x_{i,0} = 1. \tag{1.3}$$

Na forma matricial, representa-se por **Y** o vector $n \times 1$ das observações da variável dependente, por **X** a matriz $n \times (k+1)$ cuja coluna j é constituída pelas observações da variável independente x_j, por **b** o vetor $(k+1) \times 1$ dos coeficientes de regressão linear a estimar, e por $\boldsymbol{\varepsilon}$ o correspondente vetor $n \times 1$ dos termos de erro. À matriz **X** é usual chamar-se de Matriz de Planeamento.

A expressão (1.1) representa uma relação linear múltipla entre $k+1$ variáveis. Quando $k = 1$ a expressão passa a representar uma regressão linear simples.

Embora o modelo de regressão linear múltipla possa parecer, numa primeira apreciação, de âmbito de aplicação limitado, é importante reparar que as variáveis independentes podem ser incluídas após sofrerem uma transformação, por exemplo, logarítmica, exponencial, potencia, entre outras, o que faz com que o modelo seja mais flexível e abrangente. Em particular as variáveis independentes também podem ser do tipo qualitativo e serem incluídas no modelo utilizando variáveis indicatrizes, também denominadas de variáveis *dummy*.

Se a equação estimada se ajustar perfeitamente aos dados disponíveis, cada valor observado estaria sobre a reta descrita analiticamente pela equação. Contudo, regra geral, esta situação não acontece, o que faz com que se obtenha uma diferença entre os valores observados e os valores estimados da variável dependente. Esta diferença corresponde aos termos de erro, ou resíduos, que têm em conta todos os restantes fatores que não estão presentes na equação estimada.

1.1. ESTIMADORES DE MÍNIMOS QUADRADOS

O método mais usual para determinar os estimadores dos coeficientes de uma regressão linear é o **Método de Mínimos Quadrados.** Para além de ser um método relativamente simples de compreender, este permite estimar os coeficientes b_j que produzem o subespaço linear gerado pelas variáveis independentes tal que a soma dos quadrados dos resíduos seja mínima. Assim, pretende-se minimizar a seguinte soma de quadrados:

$$SQ = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} b_j x_{i,j} \right)^2 . \tag{1.4}$$

Como se pretendem os valores de b_j que minimizam SQ, é necessário resolver o seguinte sistema de $k+1$ equações a $k+1$ incógnitas:

$$\frac{\partial\, SQ}{\partial\, b_w} = 0 \tag{1.5}$$

$$\Leftrightarrow \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{k} b_j x_{i,j} \right) x_{i,w} = 0 \tag{1.6}$$

$$\Leftrightarrow \sum_{i=1}^{n} y_i x_{i,k} = \sum_{j=0}^{k} b_j \sum_{i=1}^{n} x_{i,j} x_{i,w} \text{ , com } w = 0, \dots k. \tag{1.7}$$

É possível escrever este sistema na forma matricial da seguinte forma:

$$\mathbf{X^T Y} = (\mathbf{X^T X})\mathbf{b}. \tag{1.8}$$

Assim, se a matriz $\mathbf{X^T X}$ for invertível, o vetor dos estimadores de mínimos quadrados dos coeficientes da regressão linear é dado por:

$$\hat{\mathbf{b}} = (\mathbf{X^T X})^{-1}(\mathbf{X^T Y}) \tag{1.9}$$

É importante referir que se a matriz $\mathbf{X^T X}$ não for invertível significa que existem variáveis independentes que são combinação linear de outras, pelo que estas devem ser retiradas da matriz de planeamento.

1.2. AVALIAÇÃO DO AJUSTAMENTO DO MODELO DE REGRESSÃO LINEAR

Uma vez encontradas as estimativas dos coeficientes de regressão linear, podem ser calculados os valores ajustados $\hat{y}_i$, que são dados por:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i,1} + \cdots + \hat{b}_k x_{i,k}\ , i = 1, \ldots, n. \tag{1.10}$$

Assim, as estimativas para os erros aleatórios, também chamados de resíduos, são dadas pela diferença:

$$\varepsilon_i = y_i - \hat{y}_i. \tag{1.11}$$

Tem-se então que o estimador centrado para a variância dos erros é dado por:

$$S^2 = \frac{1}{n - k' - 1} \sum_{i=1}^{n} \varepsilon_i^2, \tag{1.12}$$

onde k' é o número de coeficientes estimados.

Este estimador é também um indicador do bom ajustamento do modelo, uma vez que, quanto menor o seu valor mais importante se torna o modelo de regressão linear para explicar a variabilidade das observações.

Um outro indicador do ajustamento do modelo baseia-se na quantidade:

$$\mathrm{SQ_{TOT}} = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{1.13}$$

Esta soma, Soma de Quadrados Total, exprime a variabilidade do conjunto de observações em torno da sua média e pode ser decomposta da seguinte forma:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \left((y_i - \hat{y}_i) - (\hat{y}_i - \bar{y})\right)^2 \tag{1.14}$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{1.15}$$

Através da expressão (1.6) pode mostrar-se que:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0 \tag{1.16}$$

e desta forma pode então decompor-se a soma dos quadrados dos desvios à média dos y's da seguinte forma:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2, \tag{1.17}$$

abreviadamente, $\mathrm{SQ_{TOT}} = \mathrm{SQ_{Reg}} + \mathrm{SQ_e}$, que equivale a dizer que a variabilidade total da amostra ($\mathrm{SQ_{TOT}}$) pode decompor-se na soma de quadrados residual ($\mathrm{SQ_e}$) e na soma de quadrados devida à regressão $\left(\mathrm{SQ_{Reg}}\right)$, que refletem, respetivamente, a variabilidade devida aos erros aleatórios e a variabilidade devida ao facto de as observações seguirem um modelo de regressão linear.

Deste modo, um bom indicador do ajustamento do modelo é o coeficiente de determinação múltipla que se define como:

$$R^2 = \frac{\mathrm{SQ_{Reg}}}{\mathrm{SQ_{TOT}}} = 1 - \frac{\mathrm{SQ_e}}{\mathrm{SQ_{TOT}}}. \tag{1.18}$$

O coeficiente R^2 toma valores entre 0 e 1 e representa a percentagem de variação da amostra que é explicada pelo modelo de regressão. Assim, quanto mais próximo estiver da unidade, melhor o ajustamento do modelo, pois significa que praticamente toda a variabilidade da amostra é devida à regressão e não ao erro que se comete ao ajustar uma função linear ao valor médio das observações da amostra.

Quando o modelo de regressão não inclui termo constante, ou seja, se $b_0 = 0$, a decomposição da soma de quadrados na qual se baseia a definição de R^2 já não é válida e este coeficiente toma a forma específica:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} {y_i}^2} = \frac{\sum_{i=1}^{n} {\hat{y}_i}^2}{\sum_{i=1}^{n} {y_i}^2} \tag{1.19}$$

Apesar de, regra geral, se pretender um modelo com um valor de R^2 o mais elevado possível, tem que haver um compromisso entre este parâmetro e o número de variáveis, uma vez que ao retirar variáveis menos significativas, o valor de R^2 diminui ainda que, por vezes, muito moderadamente. Por outro lado, é importante não incluir demasiadas variáveis uma vez que a amplitude dos intervalos de previsão tente a aumentar com o número de variáveis. Adicionalmente, a inclusão de muitas variáveis no modelo, levam com frequência, à introdução de variáveis que provocam multicolinearidade[1], que por sua vez contribui para o aumento do erro de estimação e, consequentemente, da amplitude dos intervalos de previsão.

Pode ainda demonstrar-se um resultado muito útil que permite concluir que para um modelo de linear $\mathbf{Y} = \mathbf{Xb} + \boldsymbol{\varepsilon}$, em que $\boldsymbol{\varepsilon} = [\varepsilon_1 \quad \cdots \quad \varepsilon_n]^{\mathrm{T}}$ é um vetor de variáveis aleatórias i.i.d.[2] com distribuição $\mathrm{N}(0, \sigma^2)$. Nestas condições,

1. O Estimador de Mínimos Quadrados do vetor de parâmetros $\mathbf{b}$, isto é, $\hat{\mathbf{b}} = (\mathbf{X^T X})^{-1}(\mathbf{X^T Y})$, tem distribuição multinormal $\mathrm{N}\left(\mathbf{b}, \sigma^2(\mathbf{X^T X})^{-1}\right)$;

2. A variável aleatória

$$\frac{(n-k')S^2}{\sigma^2} = \frac{\boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}}{\sigma^2} \tag{1.20}$$

tem distribuição qui-quadrado com $n - k'$ graus de liberdade;

3. $\hat{\mathbf{b}}$ e S^2 são independentes.

Com este resultado, pode construir-se a estatística de teste para a hipótese de que todos os coeficientes da regressão linear são nulos, exceto o do termo constante, com o objetivo de averiguar a eficiência do modelo:

$$\mathrm{H_0}: \forall j = 1, \ldots, k, b_j = 0 \text{ vs } \mathrm{H_1}: b_j \neq 0, j = 1, \ldots, k, \tag{1.21}$$

donde, estatística de teste é dada por:

[1] O tema da multicolinearidade é abordado no capítulo 3.
[2] i.i.d.: independentes e identicamente distribuídas.

$$F = \frac{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,}{k' - 1}}{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n - k'}} = \frac{\frac{\mathrm{SQ_{Reg}}}{k' - 1}}{\frac{\mathrm{SQ_e}}{n - k'}}, \tag{1.22}$$

que sob validade da hipótese nula tem distribuição *F* com $k' - 1$ graus de liberdade no numerador e $n - k'$ graus de liberdade no denominador.

Valores muito elevados da estatística de teste *F* indicam que as variáveis independentes são importantes para explicar a variabilidade das observações. Contudo, não significa que o modelo não possa ser melhorado juntando mais variáveis ou transformando algumas das já incluídas.

De forma semelhante, e por forma a avaliar a significância de cada variável individualmente, calcula-se o seu *p-value*. Para esta análise pretende-se testar as seguintes hipóteses:

$$\mathrm{H_0}: b_j = 0 \text{ vs } \mathrm{H_1}: b_j \neq 0, j = 1, \dots k. \tag{1.23}$$

A estatística de teste usada para testar este tipo de hipóteses, em que um coeficiente assume um valor constante igual ou diferente de zero é a que se segue:

$$\mathbf{ET_j} = \frac{\hat{b}_j}{\hat{\sigma}(\hat{b}_j)}, j = 1, \dots k. \tag{1.24}$$

Na regressão linear a estatística de teste definida em (1.24) segue uma distribuição T-Student. Desta forma, para o teste definido em (1.23), tem-se a seguinte região de rejeição bilateral:

$$\Leftrightarrow \frac{|\hat{b}_j|}{\hat{\sigma}(\hat{b}_j)} > t_{n-k'-1}^{1-\alpha/2} \tag{1.25}$$

onde $\hat{\sigma}(\hat{b}_j)$ é o estimador para o desvio-padrão de $\hat{b}_j$ calculado pelo método dos mínimos quadrados e $t_{n-k'-1}^{1-\alpha/2}$ representa o quantil de ordem $1 - \alpha/2$ da distribuição T-Student com $n - k' - 1$ graus de liberdade (sendo k' o número de parâmetros estimados na regressão).

Uma vez conhecida a região de rejeição da estatística de teste, o *p-value* é o menor nível de significância (α) para o qual a hipótese nula é rejeitada, ou seja, quando *p-value* for inferior ao nível de significância escolhido, deve rejeitar-se a hipótese nula, concluindo-se que a variável é significativa.

2. CARACTERÍSTICAS AMOSTRAIS

Antes de se começar a desenvolver um modelo é importante garantir a qualidade das amostras a utilizar. Enumeram-se de seguida algumas das características que se entendem ser pertinentes para despistar possíveis problemas de integridade das observações:

- **Número de registos;** número de linhas da amostra selecionada;

- **Número de observações** (n)**:** é o número de registos preenchidos, isto é, registos que contém algum tipo de informação;

- **Percentagem de *missings:*** é a percentagem de registos que não contém informação;

- **Média:** sejam, $x_{i,j}$, com $i = 1, \dots, n$ as n observações da variável j,

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}\,; \tag{2.1}$$

- **Variância**;

$$var(x_j) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i,j} - \bar{x}_j\right)^2; \tag{2.2}$$

Note-te que é feita a divisão por $n-1$, pois pretende-se obter um estimador centrado para a variância amostral.

- **Desvio-Padrão:** é a raiz quadrada da variância,

$$s(x_j) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i,j} - \bar{x}_j\right)^2} \tag{2.3}$$

- **Mediana:** valor que se encontra a meio de uma amostra ordenada.

- **Mínimo:** menor valor observado;

- **Máximo:** maior valor observado**;**

3. DETEÇÃO DE MULTICOLINEARIDADE

A existência de multicolinearidade é um obstáculo com o qual se é confrontando aquando do ajustamento de modelos de regressão. Contudo, a existência de multicolinearidade deve ser interpretada com algum cuidado, pois a existência de correlação entre variáveis não implica necessariamente a existência de causa-efeito.

Nem sempre é fácil medir o impacto da correlação de variáveis na robustez de um modelo. Efetivamente, a existência de multicolinearidade tem implicações ao nível da variância dos coeficientes estimados tornando os estimadores bastante sensíveis a pequenas alterações do modelo. Desta forma, os coeficientes estimados tornam-se difíceis de interpretar, dificultando a escolha do melhor modelo.

Contudo, existem diversas formas que permitem detetar a presença de multicolinearidade e analisar o nível de correlação entre as variáveis. Este documento incidirá sobre um método de deteção de existência de multicolinearidade e um método de eliminação das variáveis causadoras de multicolinearidade.

3.1. MÉTODO DAS COMPONENTES DA VARIÂNCIA

Este método, tal como o seu nome indica, permite avaliar o impacto que cada variável tem sobre os coeficientes da regressão, $\hat{\mathbf{b}}$. Pelo método que se descreve abaixo, esta avaliação é feita através da análise da matriz dos vetores próprios da matriz $\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)}$, e desta forma pretende-se então avaliar o efeito da multicolinearidade sobre os coeficientes da regressão, $\hat{\mathbf{b}}_{(s)}$.

A matriz $\mathbf{X}_{(s)}$ corresponde à matriz de planeamento $\mathbf{X}$ em que as variáveis independentes são todas reduzidas à mesma escala, de modo a que uma variável cujos elementos estão próximo de zero não seja confundida com o problema da multicolinearidade. Assim, a matriz $\mathbf{X}_{(s)}$ pode ser escrita da seguinte forma:

$$\mathbf{X}_{(s)} = \mathbf{X}\mathbf{D}_{(s)}^{-1}, \tag{3.1}$$

com:

$$\mathbf{D}_{(s)} = \begin{bmatrix} \|\mathrm{x}_0\| & 0 & \cdots & 0 \\ 0 & \|\mathrm{x}_1\| & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \|\mathrm{x}_k\| \end{bmatrix} \text{ e } \hat{\mathbf{b}}_{(s)} = \mathbf{D}_{(s)}\hat{\mathbf{b}}, \tag{3.2}$$

onde, $\|\mathrm{x}_j\|$ é a norma do vetor cujas entradas são as observações da variável x_j:

$$\|\mathrm{x}_j\| = \sqrt{x_{1,j}^2 + \cdots + x_{n,j}^2}, j = 0, \ldots, k. \tag{3.3}$$

Como a matriz $\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)}$ é uma matriz simétrica, pode ainda ser escrita da seguinte forma:

$$\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)} = \mathbf{\Gamma}\mathbf{D}_{\lambda}\mathbf{\Gamma}^{\mathrm{T}}, \tag{3.4}$$

em que

$$\mathbf{D}_{\lambda} = \begin{bmatrix} \lambda_0 & 0 & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix}, \tag{3.5}$$

onde λ_j são os valores próprios de $\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)}$, e

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma_{1,0} & \gamma_{1,1} & \cdots & \gamma_{1,k} \\ \gamma_{2,0} & \gamma_{2,1} & \cdots & \gamma_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{k,0} & \gamma_{k,1} & \cdots & \gamma_{k,k} \end{bmatrix}, \tag{3.6}$$

é a matriz ortogonal cujas colunas são os vetores próprios de $\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)}$.

Como a matriz de covariâncias de $\hat{\mathbf{b}}_{(s)}$ é dada por:

$$cov\left(\hat{\mathbf{b}}_{(s)}\right) = \sigma^2\left(\mathbf{X}_{(s)}^{\mathrm{T}}\mathbf{X}_{(s)}\right)^{-1}, \tag{3.7}$$ [3]

pela equação (3.4) tem-se que:

$$cov\left(\hat{\mathbf{b}}_{(s)}\right) = \sigma^2\left(\mathbf{\Gamma}\mathbf{D}_{\lambda}\mathbf{\Gamma}^{\mathrm{T}}\right)^{-1} = \sigma^2\mathbf{\Gamma}\mathbf{D}_{\boldsymbol{\lambda}}^{-\mathbf{1}}\mathbf{\Gamma}^{\mathrm{T}}. \tag{3.8}$$

Assim, pode escrever-se a variância de cada elemento do vetor de estimadores de mínimos quadrados dos coeficientes da regressão como função dos vetores próprios e valores próprios da matriz de covariâncias, ou seja:

$$var\left(\hat{b}_{j\,(s)}\right) = \sigma^2 \sum_{w=0}^{k} \lambda_w^{-1}\gamma_{j,w}^2\,, j = 0, \ldots, k. \tag{3.9}$$

Os valores $\lambda_w^{-1}\gamma_{j,w}^2$, são chamados de **componentes da variância** de $\hat{b}_{j\,(s)}$ e os coeficientes

$$\phi_{w,j} = \lambda_w^{-1}\gamma_{j,w}^2 \Big/ \sum_{l=0}^{k} \lambda_w^{-1}\gamma_{j,l}^2 \tag{3.10}$$

correspondem à proporção da variância do j-ésimo coeficiente de regressão, $\hat{b}_{j\,(s)}$, que é explicada pelo w-ésimo valor próprio.

Como a cada valor próprio próximo de zero corresponde apenas uma única equação linear entre variáveis, a análise das **componentes da variância** para cada um desses valores próprios permite identificar quais as variáveis envolvidas em cada uma dessas relações. Assim, para um determinado λ_j próximo de zero, as componentes da variância $\phi_{w,j}$ próximas da unidade indicam as variáveis X_j envolvidas nessa relação.

Adicionalmente, uma vez obtidos os valores próprios, pode ser avaliada a grandeza dos valores próprios relativamente a outros através dos números condição que são definidos por:

$$\eta_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}}, j = 0, \ldots, k, \tag{3.11}$$

onde λ_{max} é o maior dos valores próprios. Em geral, um valor próprio para o qual $\eta_j > 30$ indica que existe uma relação linear entre variáveis da matriz de planeamento. Devem então ser identificadas as variáveis que provocam um aumento exagerado da variância nos estimadores dos coeficientes da regressão. Desta forma, é usual construir a Tabela 3.1.

[3] 2ª condição de Gauss-Markov: $var(\varepsilon_i) = \sigma^2$

Tabela 3.1 – Tabela de proporções da variância dos coeficientes de regressão

Valores Próprios	**Números condição**	**Proporções de**			
		$\boldsymbol{var(\hat{b}_0)}$	$\boldsymbol{var(\hat{b}_1)}$	...	$\boldsymbol{var(\hat{b}_0)}$
λ_0	η_0	$\frac{\sigma^2\gamma_{1,0}^2}{\lambda_0 var(\hat{b}_0)}$	$\frac{\sigma^2\gamma_{1,1}^2}{\lambda_0 var(\hat{b}_1)}$	...	$\frac{\sigma^2\gamma_{1,k}^2}{\lambda_0 var(\hat{b}_k)}$
λ_1	η_1	$\frac{\sigma^2\gamma_{2,0}^2}{\lambda_1 var(\hat{b}_0)}$	$\frac{\sigma^2\gamma_{2,1}^2}{\lambda_1 var(\hat{b}_1)}$	...	$\frac{\sigma^2\gamma_{2,k}^2}{\lambda_1 var(\hat{b}_k)}$
...	...	...	...	...	...
λ_k	η_k	$\frac{\sigma^2\gamma_{n,0}^2}{\lambda_k var(\hat{b}_0)}$	$\frac{\sigma^2\gamma_{n,1}^2}{\lambda_k var(\hat{b}_1)}$	...	$\frac{\sigma^2\gamma_{n,k}^2}{\lambda_k var(\hat{b}_k)}$

A análise desta tabela consiste, essencialmente, em identificar os valores próprios próximos de zero e, na linha correspondente, procurar os $\phi_{w,j}$ próximos da unidade. Os elementos em cada coluna da tabela têm soma igual à unidade. Note-se ainda que, se uma variável estiver envolvida em mais de uma relação linear a sua contribuição subdivide-se pelas linhas correspondentes aos valores próprios associados e desta forma, mesmo os valores de $\phi_{w,j}$ que não muito elevados podem ser indicadores de multicolinearidade.

3.2. MÉTODO DOS FATORES DE INFLAÇÃO DAS VARIÂNCIAS

Uma forma mais simples de detetar o grau de dependência entre cada variável independente x_j e as restantes variáveis, é examinar o valor de R_j^2, que é o valor de R^2 quando se faz a regressão de x_j em função das restantes variáveis:

$$x_j = b_1 x_1 + \cdots + b_{j-1} x_{j-1} + b_{j+1} x_{j+1} + \cdots + b_k x_k + \varepsilon_j, \tag{3.12}$$

Note-se que para este método é usada a matriz $\mathbf{X}_{(r)}$ que é simplesmente a matriz de planeamento sem a coluna do termo constante, e é representada da seguinte forma:

$$\mathbf{X}_{(r)} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,1} & \cdots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{bmatrix} \tag{3.13}$$

A tolerância da variável x_j (TOL_j) define-se da seguinte forma:

$$TOL_j = 1 - R_j^2. \tag{3.14}$$

Como R_j^2 toma sempre valores entre 0 e 1, se $TOL_j \approx 1$ significa que a variável x_j tem fraca relação com as restantes variáveis, enquanto que se $TOL_j \approx 0$ significa que a variável x_j tem uma relação aproximadamente linear com as restantes variáveis independentes.

Por sua vez, o **fator de inflação das variâncias** ou *Variance Inflaction Factor* (VIF_j) é dado por:

$$VIF_j = \frac{1}{TOL_j} = \frac{1}{1 - R_j^2}. \tag{3.15}$$

É fácil concluir que um valor de VIF_j próximo da unidade significa que não existe dependência entre x_j e as restantes variáveis, enquanto que valores de VIF_j muito grandes indicam a presença de multicolinearidade.

Uma das formas mais usadas para a deteção e eliminação de variáveis causadoras de multicolinearidade é através da análise da matriz inversa da matriz de correlações $(\mathcal{R}^{-1})$ da matriz $\mathbf{X}_{(r)}$.

No entanto, pode ser demostrado um resultado que diz que os elementos da diagonal da matriz $\mathcal{R}^{-1}$ são os fatores de inflação das variâncias. Desta forma, ao serem retiradas as variáveis correspondes a entradas de valores muito elevados na diagonal da matriz $\mathcal{R}^{-1}$ estão a ser retiradas as variáveis com maior efeito de multicolinearidade.

A matriz de correlações $\mathcal{R}$ define-se da seguinte forma:

$$\mathcal{R} = \begin{bmatrix} 1 & correl(\mathrm{x}_1, \mathrm{x}_2) & \cdots & correl(\mathrm{x}_1, \mathrm{x}_k) \\ correl(\mathrm{x}_2, \mathrm{x}_1) & 1 & \cdots & correl(\mathrm{x}_2, \mathrm{x}_k) \\ \vdots & \vdots & \ddots & \vdots \\ correl(\mathrm{x}_k, \mathrm{x}_1) & correl(\mathrm{x}_k, \mathrm{x}_2) & \cdots & 1 \end{bmatrix}, \tag{3.16}$$

onde,

$$correl(\mathrm{x}_l, \mathrm{x}_m) = \frac{\sum_{i=1}^{n} \left(\mathrm{x}_{i,l} - \bar{\mathrm{x}}_l\right)\left(\mathrm{x}_{i,m} - \bar{\mathrm{x}}_m\right)}{\sqrt{\sum_{i=1}^{n} \left(\mathrm{x}_{i,l} - \bar{\mathrm{x}}_l\right)^2 \sum_{i=1}^{n} \left(\mathrm{x}_{i,m} - \bar{\mathrm{x}}_m\right)^2}} \tag{3.17}$$

É importante referir que esta análise apenas revela as relações existentes entre cada par de variáveis.

Assim, este método pode ser resumido nos seguintes passos:

Passo 1: Definir um valor de R^2 a partir do qual deixa de ser aceitável aceitável qualquer relação linear que possa existir entre as variáveis independentes;

Passo 2: Calcular a matriz $\mathcal{R}^{-1}$;

Passo 3: Se existir algum elemento na diagonal da matriz $\mathcal{R}^{-1}$ com valor superior ao VIF, retira-se a variável correspondente ao maior desses elementos;

Passo 4: Repete-se o processo desde o passo 2 até que todos os elementos da diagonal da matriz $\mathcal{R}^{-1}$ sejam inferiores ou iguais ao VIF.

Como este método testa a relação entre pares de variáveis independentes, e pretende-se reduzir a multicolinearidade, não é desejável obter valores de R^2 muito elevados para cada par testado. Assim, considera-se teoricamente aceitável uma relação linear para valores de VIF_j inferiores ou iguais a 2,5, isto é, é aceitável qualquer relação linear entre duas variáveis independentes desde que $R_j^2 \leq 0{,}6(6)$.

4. MODELOS DE REGRESSÃO LOGÍSTICA

No primeiro capítulo foram abordados os modelos de regressão simples e múltipla como uma forma de exprimir e prever uma quantidade como uma função linear de uma ou mais variáveis. Neste capítulo a atenção recairá sobre o mesmo tema mas com o objetivo de dar resposta aos casos em que a variável dependente é binária, ou seja, só toma valor "1" ou "0", consoante um determinado acontecimento se verifica ou não. Assim, assume-se que a variável dependente, Y, tem distribuição Bernoulli, ou seja,:

$$\mathrm{P}(\mathrm{Y}=1)=p \text{ e } \mathrm{P}(\mathrm{Y}=0)=1-p \tag{4.1}$$

No entanto, num modelo de regressão binária a probabilidade de sucesso depende dos valores das variáveis dependentes, ou seja:

$$p=\mathrm{p}(\mathrm{x}_1,\dots,\mathrm{x}_k)=\mathrm{P}(\mathrm{Y}=1) \tag{4.2}$$

Em geral, a função $\mathrm{p}(\mathrm{x}_1,\dots,\mathrm{x}_k)$ pertence a uma família paramétrica de curvas, com propriedades especiais que as tornam particularmente útil no ajustamento deste tipo de modelos.

Uma das funções mais utilizadas para descrever a variação da probabilidade em termos da variação das variáveis independentes é a função logística, que é definida da seguinte forma:

$$\mathrm{p}(\mathrm{x}_1,\dots,\mathrm{x}_k;\ b_0,b_1,\dots,b_k)=\frac{e^{(b_0+b_1\mathrm{x}_1+\cdots+b_k\mathrm{x}_k)}}{1+e^{(b_0+b_1\mathrm{x}_1+\cdots+b_k\mathrm{x}_k)}}=\frac{1}{1+e^{-(b_0+b_1\mathrm{x}_1+\cdots+b_k\mathrm{x}_k)}}. \tag{4.3}$$

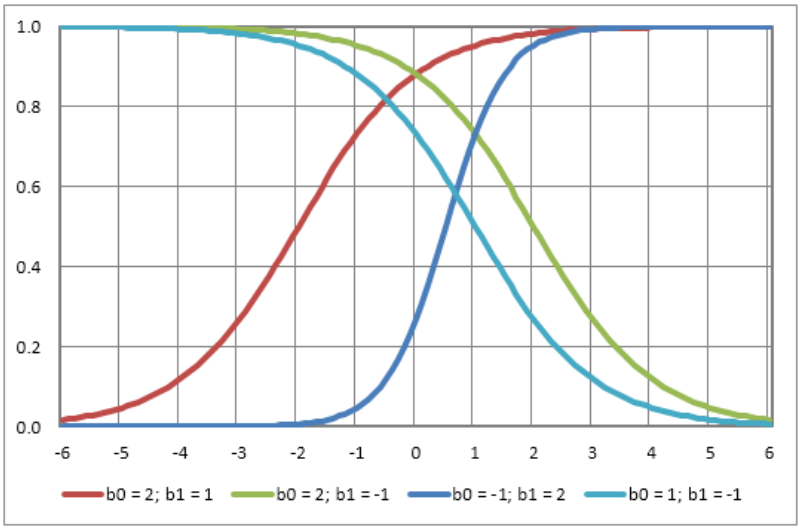

Figura 4.1 – Função Logística: $y=e^{b_0+b_1x_1}/1+e^{b_0+b_1x_1}$

Assim, para um conjunto de observações da variável dependente, isto é, para um vetor **Y** em que cada elemento corresponde a uma observação $y_i, i = 1, \dots, n$ vem que:

$$\mathrm{p}(y_i = 1) = \mathrm{p}\big(y_i = 1 | x_{i,1}, \dots, x_{i,k}\big) = \frac{e^{(\mathrm{x}_i^{\mathrm{T}}\mathbf{b})}}{1 + e^{(\mathrm{x}_i^{\mathrm{T}}\mathbf{b})}}, \tag{4.4}$$

em que $\mathrm{x}_i^{\mathrm{T}} = [x_{i,1} \quad \cdots \quad x_{i,k}]$ é o vetor das variáveis independentes correspondentes a y_i e **b** definido da mesma forma que em (1.3).

Do mesmo modo como foi referido para a regressão linear, também na regressão logística podem ser incluídas transformações de uma mesma variável, dando assim uma maior flexibilidade ao modelo a desenvolver.

4.1. Estimadores de Máxima Verosimilhança

O método usado na estimação dos coeficientes da regressão logística é o **Método da Máxima Verosimilhança**, que se sabe produzir estimadores com boas propriedades, consiste na maximização da função de verosimilhança, e neste caso é dada por:

$$\begin{aligned} &\mathcal{L}(y_1, y_2, \dots y_n;\ b_0, b_1, \dots, b_k) \\ &= \prod_{i=1}^{n} \mathrm{p}\big(x_{i,1}, \dots, x_{i,k}\big)^{y_i} \Big(1 - \mathrm{p}\big(x_{i,1}, \dots, x_{i,k}\big)^{1-y_i}\Big), \end{aligned} \tag{4.5}$$

ou, em notação simplificada:

$$\mathcal{L} = \prod_{i=1}^{n} \mathrm{p}_i^{y_i} (1 - \mathrm{p}_i)^{1-y_i}. \tag{4.6}$$

Uma vez que a função logaritmo é uma função contínua, monótona e crescente, determinar o maximizante do logaritmo de uma função é equivalente a determinar o maximizante dessa mesma função. Desta forma, e com o objetivo de facilitar a construção da solução deste problema, é usual usar-se o logaritmo da verosimilhança, ou logverosimilhança. Desta forma, a logverosimilhança pode ser escrita como:

$$\ln(\mathcal{L}) = \sum_{i=1}^{n} y_i \ln(\mathrm{p}_i) + \sum_{i=1}^{n} (1 - y_i) \ln(1 - \mathrm{p}_i). \tag{4.7}$$

Assim, a derivada da logverosimilhança em ordem a cada um dos coeficientes$b_j, j = 0, \dots, k$ é dada por:

$$\frac{\partial \ln(\mathcal{L})}{\partial b_j} = \sum_{i=1}^{n} y_i \frac{\partial \mathrm{p}_i \big/ \partial b_j}{\mathrm{p}_i} - \sum_{i=1}^{n} (1 - y_i) \frac{\partial \mathrm{p}_i \big/ \partial b_j}{1 - \mathrm{p}_i}, \tag{4.8}$$

onde,

$$\frac{\partial \mathrm{p}_i}{\partial b_j} = x_{i,j} \mathrm{p}_i (1 - \mathrm{p}_i). \tag{4.9}$$

Como se pretendem os valores de b_j que maximizam $\ln(\mathcal{L})$, é necessário resolver o seguinte sistema de $k+1$ equações a $k+1$ incógnitas:

$$\frac{\partial \ln(\mathcal{L})}{\partial b_j} = \sum_{i=1}^{n} x_{i,j}(y_i - \mathrm{p}_i) = 0, j = 0, \dots, k \tag{4.10}$$

É relevante reparar que estas equações são muito semelhantes às equações normais usadas para a regressão linear simples. O problema é que, no caso da regressão linear, o valor médio das observações é uma função linear das variáveis independentes e, neste caso, o valor médio é a função logística, o que faz com que as equações normais deixem de ser equações lineares. Na realidade, estas não têm solução analítica, sendo necessário recorrer a métodos numéricos como, por exemplo, o método de Newton-Raphson.

Para aplicar o método de Newton-Raphson é conveniente usar a notação matricial:

$$\frac{\partial \ln(\mathcal{L})}{\partial \mathbf{b}} = \mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \mathbf{\Pi}) = 0, \tag{4.11}$$

onde **X** é a matriz de planeamento e **Y** o vetor das observações da variável dependente já definidos, e **Π** é um vetor $n \times 1$ cujas entradas são os elementos p_i.

O método de Newton-Raphson é um método iterativo em que cada iteração é definida à custa da anterior. Neste caso particular, a aplicação do método de Newton-Raphson à equação acima especificada faz com que se obtenha que a iteração w é calculada pela seguinte expressão:

$$\mathbf{b}^{(w+1)} = \mathbf{b}^{(w)} - \mathbf{H}^{-1^{(w)}} \times \left.\frac{\partial \ln(\mathcal{L})}{\partial \mathbf{b}}\right|_{\mathbf{b}^{(w)}}, \tag{4.12}$$

em que $\mathbf{H}^{-1^{(w)}}$ representa a inversa da matriz Hessiana. Esta é a matriz das segundas derivadas da logverosimilhança calculadas na iteração w,

$$\frac{\partial^2 \ln(\mathcal{L})}{\partial b_{\boldsymbol{l}} \partial b_{\boldsymbol{m}}} = \sum_{i=1}^{n} x_{i,m} \times x_{i,l} \times \mathrm{p}_i(1 - \mathrm{p}_i), \tag{4.13}$$

ou, na forma matricial:

$$\mathbf{H} = \mathbf{X}^{\mathbf{T}}\mathbf{V}\mathbf{X}, \tag{4.14}$$

em que **V** é uma matriz diagonal cujos elementos da diagonal principal são da forma $\mathrm{p}_i(1 - \mathrm{p}_i)$.

Tem-se então

$$\mathbf{b}^{(w+1)} = \mathbf{b}^{(w)} - (\mathbf{X}^{\mathbf{T}}\mathbf{V}^{(\boldsymbol{w})}\mathbf{X})^{-1} \times \mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \mathbf{\Pi}^{(\boldsymbol{w})}), \tag{4.15}$$

O processo termina quando duas iterações sucessivas produzem valores muito semelhantes, nomeadamente quando se verificar:

$$\left\|\mathbf{b}^{(k+1)} - \mathbf{b}^{(k)}\right\| < \delta, \tag{4.16}$$

em que δ é a precisão que deve ser definida à partida.

4.2. AVALIAÇÃO DO AJUSTAMENTO DO MODELO DE REGRESSÃO LOGÍSTICA

Os estimadores de máxima verosimilhança podem, em condições razoavelmente gerais, ter distribuição limite normal, com valor médio igual ao verdadeiro valor dos parâmetros e matriz de covariâncias dada pela inversa da matriz de informação de Fisher, $\mathbf{I}(\mathbf{b})$. As condições para que se verifique este resultado são dadas por um teorema de Fahrmeir e Kaufmann, que nos pode fornecer informação muito útil sobre a distribuição assintótica dos estimadores de máxima verosimilhança e, portanto, sobre a construção de testes de hipóteses sobre os coeficientes de regressão.

Sejam $y_i, i = 1, \dots n,$ variáveis aleatórias independentes com distribuição Bernoulli que seguem um modelo de regressão logística, isto é, cujo parâmetro p_i é dado por:

$$\mathrm{p}_i = \frac{e^{(b_0 + b_1 \mathrm{x}_1 + \cdots + b_k \mathrm{x}_k)}}{1 + e^{(b_0 + b_1 \mathrm{x}_1 + \cdots + b_k \mathrm{x}_k)}}. \tag{4.17}$$

Suponha-se ainda que se verificam as seguintes condições:

1. $\lambda_{min}\mathbf{I}(\mathbf{b}) = \lambda_{\boldsymbol{min}}\left(\mathbf{X^T V(b) X}\right) \to +\infty$ *quando* $n \to +\infty$;
2. $tr\left[\mathbf{X^T I^{-1}(b) X}\right] \to 0$ quando $n \to +\infty$,

em que λ_{min} é o menor valor próprio da matriz $\mathbf{I}(\mathbf{b})$. Nestas condições os estimadores de máxima verosimilhança do vetor de coeficientes, $\hat{\mathbf{b}}$, são consistentes e com distribuição assintoticamente normal, ou seja,

$$\mathbf{I}^{1/2}(\mathbf{b})\left(\hat{\mathbf{b}} - \mathbf{b}\right) \xrightarrow{\mathbf{L}} \mathrm{N}(0, \mathbf{I}). \tag{4.18}$$[4]

Uma forma de avaliar o ajustamento de um modelo de regressão logística é aplicando um teste de razão de verosimilhanças generalizado à hipótese:

$$\mathrm{H}_0: b_1 = b_2 = \cdots = b_k = 0 \text{ vs } \mathrm{H}_1: \exists\, j, j = 1, \dots, k \text{ tal que } b_j \neq 0 \tag{4.19}$$

Este teste avalia a necessidade de ajustar um modelo de regressão logística e é baseado na estatística de teste:

$$\lambda = \frac{L_0}{L_1}, \tag{4.20}$$

em que L_0 se refere à função de verosimilhança calculada nos estimadores de máxima verosimilhança sobre a validade da hipótese nula, $\mathcal{L}(y_1, y_2, \dots y_n;\ b_0)$, e L_1 é a verosimilhança calculada nos estimadores de máxima verosimilhança sem qualquer restrição, $\mathcal{L}(y_1, y_2, \dots y_n;\ b_0, b_1, \dots, b_k)$.

Desta forma, o quociente $\frac{L_0}{L_1}$ representa a perda, em termos de verosimilhança, ao não incluir as variáveis no modelo.

[4] **Convergência em Lei**: Seja $\{X_n\}_{n\geq 1}$ uma sucessão de variáveis aleatórias. Dizemos que X_n converge em lei (ou em distribuição) para Y se e só se $\lim_{n\to\infty} F_{X_n}(x) = F_Y(x)$, para todo x que seja ponto de continuidade de F_Y, e denotamos tal facto por $X_n \xrightarrow{\mathrm{L}} Y$

Pelo teorema de Wilks, tem-se ainda que $-2 \times \ln\left(\frac{L_0}{L_1}\right)$ tem distribuição aproximadamente qui-quadrado com *k* graus de liberdade. Pode então concluir-se que quanto maior for $-2 \times \ln\left(\frac{L_0}{L_1}\right)$, maior será a probabilidade de rejeitar a hipótese nula em que as variáveis não estão presentes.

À semelhança do que foi descrito para a regressão linear, também na regressão logística é relevante testar a significância de cada variável individualmente, pelo mesmo teste já referido:

$$\mathrm{H}_0: b_j = 0 \text{ vs } \mathrm{H}_1: b_j \neq 0 \tag{4.21}$$

A estatística de teste também é dada pela mesma expressão:

$$\mathbf{ET_j} = \frac{\hat{b}_j}{\hat{\sigma}\left(\hat{b}_j\right)}, j = 1, \dots k, \tag{4.22}$$

no entanto, na regressão logística, a estatística de teste definida em (4.22) segue uma distribuição Normal Padrão, em consequência do teorema já enunciado. Considerando $\hat{\sigma}^2\left(\hat{b}_j\right)$ o estimador da variância de $\hat{b}_j$, ou seja, o j-ésimo elemento da diagonal da matriz $\mathbf{I}(\mathbf{b})^{-1} = \left(\mathbf{X^T V}(\hat{\mathbf{b}})\mathbf{X}\right)^{-1}$, para o teste definido em (4.21), tem-se a seguinte região de rejeição bilateral:

$$\Leftrightarrow \frac{\left|\hat{b}_j\right|}{\hat{\sigma}\left(\hat{b}_j\right)} > q_{1-\frac{\alpha}{2}} \tag{4.23}$$

onde $q_{1-\frac{\alpha}{2}}$ representa o quantil de ordem $1 - {}^{\alpha}/_{2}$ da distribuição normal padrão. Convém notar que, neste caso, se trata de um teste aproximado, uma vez que a distribuição da estatística de teste é apenas assintoticamente normal.

Uma vez conhecida a região de rejeição da estatística de teste, o *p-value* é o menor nível de significância (α) para o qual a hipótese nula é rejeitada, ou seja, quando *p-value* for inferior ao nível de significância escolhido, deve rejeitar-se a hipótese nula, concluindo-se que a variável é significativa.

4.3. MATRIZ DE CONFUSÃO E CURVA ROC

Uma forma de avaliar a qualidade do ajustamento de uma regressão logística é comparando os valores observados com os valores previstos, ou seja, os valores obtidos pelo modelo. Na realidade, o que o modelo estima para cada observação é a probabilidade de esta tomar o valor "1" ou "0". Assim, para cada observação o valor estimado para a probabilidade de ocorrência de "sucesso" estará no intervalo $]0,1[$. Assim, tendo em consideração o valor desta probabilidade será necessário prever se a variável dependente toma o valor "0" ou "1". Para valores da probabilidade superiores a um determinado limite, ao qual se chama ponto de corte (*cut-off*), prevê-se o valor da variável como sendo "1" e para valores inferiores prevê-se como sendo "0". Desta forma, tem-se n pares de observações binários constituídos pelos valores observados da variável dependente e os respetivos valores previstos, tendo em consideração o ponto de corte definido. Isto permite avaliar a concordância entre os valores observados e os valores previstos.

É usual denominar de "valores positivos" os valores (observados e previstos) iguais à unidade, e "valores negativos" aos que são nulos. Assim podem ser definidos os seguintes conceitos:

- **Verdadeiros Positivos (VP):** valores positivos que foram previstos como tal;
- **Falsos Positivos (FP):** valores negativos que foram erradamente previstos como positivos;
- **Falsos Negativos (FN):** valores positivos que foram erradamente previstos como negativos;
- **Verdadeiros Negativos (VN):** valores negativos que foram previstos como tal.

Uma vez classificados todos os pares de valores observados e previstos, podem ser apresentados os resultados, no que usualmente se chama de **Matriz de Confusão** (*Confusion Matrix*), e que assume a forma que se apresenta na Tabela 4.1.

Tabela 4.1 - Matriz de Confusão

		Observados		
		Positivos	**Negativos**	**Total**
Previstos	**Positivos**	VP	FP (Erro de tipo II)	VP + FP
	Negativos	FN (Erro de tipo I)	VN	FN + VN
	Total	VP + FN	FP + VN	Nº Obs.

Idealmente, pretendem-se números totais de falsos positivos (FP) e de falsos negativos (FN) o mais pequenos possível. Estes valores podem ser alterados definindo diferentes pontos de corte.

Uma vez construída a matriz de confusão, podem ser definidas as seguintes probabilidades:

- **Sensibilidade** (***S***)**:** proporção de verdadeiros positivos:

$$S = \frac{VP}{VP + FN}; \tag{4.24}$$

- **Especificidade** (***E***)**:** proporção de verdadeiros negativos:

$$E = \frac{VN}{VN + FP}; \tag{4.25}$$

- (**1-S):** proporção de falsos negativos:

$$1 - S = \frac{FN}{VP + FN}; \tag{4.26}$$

- (**1-E)**: proporção de falsos positivos:

$$1 - E = \frac{FP}{VN + FP}; \tag{4.27}$$

- **Precisão**: proporção de positivos e negativos previstos corretamente:

$$\text{Precisão} = \frac{VP + VN}{n}; \tag{4.28}$$

- **Taxa de Erro**: proporção de positivos e negativos previstos erradamente:

$$\text{Taxa de erro} = \frac{FP + FN}{n} = 1 - \text{Precisão}; \tag{4.29}$$

Uma vez consolidados estes conceitos, é fácil entender que um bom modelo será aquele em que tanto a proporção de falsos positivos $(1 - E)$ como a proporção de falsos negativos $(1 - S)$ são baixas, não esquecendo que estes valores estão dependentes do ponto de corte que deverá ser definido *à priori*. Contudo, limitar a proporção de falsos negativos ou de falsos positivos tem a ver com a natureza de cada problema e com a gravidade das consequências de cada tipo de erro (Tipo I e Tipo II)

Uma forma de procurar o melhor ponto de corte será representando a **Curva ROC** (*Receiving Operating Characteristics*), que para cada valor de ponto de corte, reflete a percentagem de verdadeiros positivos (S) contra a percentagem de falsos positivos $(1 - E)$, permitindo desta forma visualizar a performance de uma variável dependente binária.

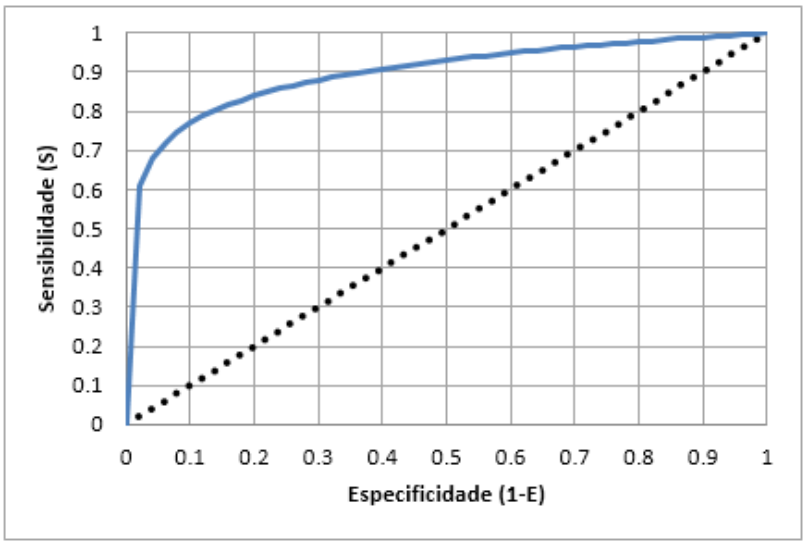

Figura 4.2 – Exemplo de uma curva ROC

5. MÉTODOS DE SELEÇÃO DE VARIÁVEIS

A parte mais relevante de qualquer modelo preditivo é saber quais as variáveis que melhor explicam determinado acontecimento ou característica. Quando se inicia a construção de um modelo, diversas variáveis podem surgir como sendo potencialmente explicativas da variável dependente. Contudo, apesar de existir a tendência de incluir determinada variável por se considerar, da experiência de cada um, que poderá será explicativa da variável dependente, tal relação pode não existir do ponto de vista estatístico e tendo em conta a amostra disponível. Adicionalmente, o analista pode querer incluir no seu modelo variáveis relacionadas entre si que podem levar a um enviesamento dos resultados finais.

Desta forma, antes de proceder ao desenvolvimento dos modelos, é necessário efetuar uma análise preliminar das variáveis de modo a excluir variáveis dependentes que possam vir a condicionar negativamente os modelos desenvolvidos, por exemplo, por provocarem problemas de multicolinearidade ou terem uma dispersão demasiado reduzida.

5.1. MÉTODO DE SELEÇÃO REGRESSIVA

O método de seleção regressiva, ou *Backward Elimination*, tal como o seu nome indica, é um método de exclusão de variáveis. Em cada iteração é efetuado um teste estatístico para identificar qual a variável independente menos significativa, ou seja, qual a variável independente com menor relação com a variável dependente. Uma vez identificada a variável menos significativa, o processo deve ser repetido – sem a variável identificada na iteração anterior – até que todas as variáveis que permanecem no modelo sejam significativas.

Desta forma, o processo pode ser resumido nos seguintes passos:

Passo 1: Começar com todas as variáveis;

Passo 2: Calcular o *p-value* do modelo com as variáveis selecionadas;

Passo 3: Identificar a variável com maior *p-value*;

Passo 4.1: Se o maior *p-value* for superior ou igual ao nível de significância escolhido, deve retirar-se essa variável e repetir-se o processo desde o passo 2;

Passo 4.2: Se o maior *p-value* obtido for inferior ao nível de significância escolhido significa que todas as variáveis testadas são significativas e o processo termina.

5.2. MÉTODO DE SELEÇÃO PROGRESSIVA

O método de seleção progressiva, ou *Forward Selection*, tal como o seu nome indica, é um método de inclusão de variáveis. Em cada iteração pretende-se testar a inclusão de cada variável juntamente com as varáveis previamente selecionadas, a fim de identificar a variável com maior significância de entre as restantes. Em seguida deve ser testada a significância dessa variável de forma a decidir se deve ou não ser incluída no modelo. Este processo deve ser repetido até que exista alguma variável que deixa de ser significativa.

Desta forma, o processo pode ser resumido nos seguintes passos:

Passo 1: Começar com nenhuma variável;

Passo 2: Calcular o R^2 ou a Razão de Verosimilhanças (caso se trate de um modelo de regressão linear ou logística, respetivamente) dos modelos com cada uma das variáveis dependentes restantes, individualmente;

Passo 3: Escolher a variável mais significativa: pretende-se continuar o processo com a variável que provoque um maior aumento de R^2 ou uma maior diminuição da Razão de Verosimilhanças;

Passo 4: Calcular o *p-value* da variável selecionada;

Passo 4.1: Se o *p-value* for inferior ao nível de significância escolhido, a variável é significativa, deve ser incluída no modelo e repete-se o processo desde o passo 2 com as restantes variáveis;

Passo 4.2: Se o maior *p-value* obtido for superior ou igual ao nível de significância escolhido significa que a variável não é significativa e o processo termina.

5.3. MÉTODO DE SELEÇÃO *STEPWISE*

Ao contrário dos métodos anteriormente descritos, o método de seleção *Stepwise* não segue uma única direção no que toca à seleção das variáveis mais explicativas da variável independente. Por sua vez, este método consiste na inclusão e remoção sequencial de variáveis independentes, até que não existam mais variáveis significativas a incluir no modelo, ou que todas as variáveis já incluídas no modelo sejam significativas.

A fim de evitar que as variáveis estejam consecutivamente a entrar e a sair do modelo, é recomendado testar com níveis de significância distintos os dois passos que mais se destacam neste processo. Mais concretamente, deve ser-se mais permissivo com a entrada de variáveis no modelo, do que com a sua saída, isto é, o nível de significância para incluir uma variável deverá ser mais baixo do que o nível de significância considerado para colocar uma nova variável no modelo.

Desta forma, o processo pode ser resumido nos seguintes passos:

Passo 1: Começar com nenhuma variável;

Passo 2: Calcular o R^2 ou a Razão de Verosimilhanças (caso se trate de um modelo de regressão linear ou logística, respetivamente) dos modelos com cada uma das variáveis dependentes restantes, individualmente;

Passo 3: Escolher a variável mais significativa: pretende-se continuar o processo com a variável que provoque um maior aumento de R^2 ou uma maior diminuição da Razão de Verosimilhanças;

Passo 4: Calcular o *p-value* da variável selecionada;

Passo 4.1: Se o maior *p-value* obtido for superior ou igual ao nível de significância escolhido para a entrada de variáveis significa que a variável não é significativa e o processo termina;

Passo 4.2: Se o *p-value* for inferior ao nível de significância escolhido para a entrada de variáveis, a variável é significativa, deve ser incluída no modelo e continua-se para o passo 5;

Passo 5: Testam-se as restantes variáveis já incluídas no modelo para averiguar a existência de variáveis não significativas;

Passo 5.1: Em caso positivo, todas as variáveis cujo *p-value* seja superior ou igual ao nível de significância escolhido para a saída de variáveis, devem ser retiradas e deve repetir-se o processo desde o passo 2;

Passo 5.2: Em caso negativo, todas as variáveis são significativas, e deve continuar-se o processo desde o passo 2.

2ª PARTE – IMPLEMENTAÇÃO E RESULTADOS

6. IMPLEMENTAÇÃO

Neste capítulo apresenta-se a forma de utilização dos programas em VBA que permitem aplicar os modelos e procedimentos que foram descritos na primeira parte deste relatório. Nesse sentido, neste capítulo irão ser apresentados os menus dos programas elaborados, realçando os pontos que se entendem como sendo mais relevantes para o utilizador.

Repare-se que, para todos os procedimentos implementados, as diversas variáveis devem estar dispostas em coluna numa folha Excel, e o utilizador deve sempre selecionar a sua amostra incluindo na 1ª linha o nome das respetivas variáveis.

É ainda relevante notar que os programas desenvolvidos devem ser utilizados com cuidado caso as capacidades do computador sejam mais limitadas, pois os procedimentos podem fazer com que o Microsoft Office Excel bloqueie quando são selecionadas amostras com mais de 20 variáveis.

Os códigos dos programas desenvolvidos podem ser consultados na íntegra no anexo deste documento.

Para aceder aos procedimentos desenvolvidos foram criados cinco botões conforme se pode ver na Figura 6.1.

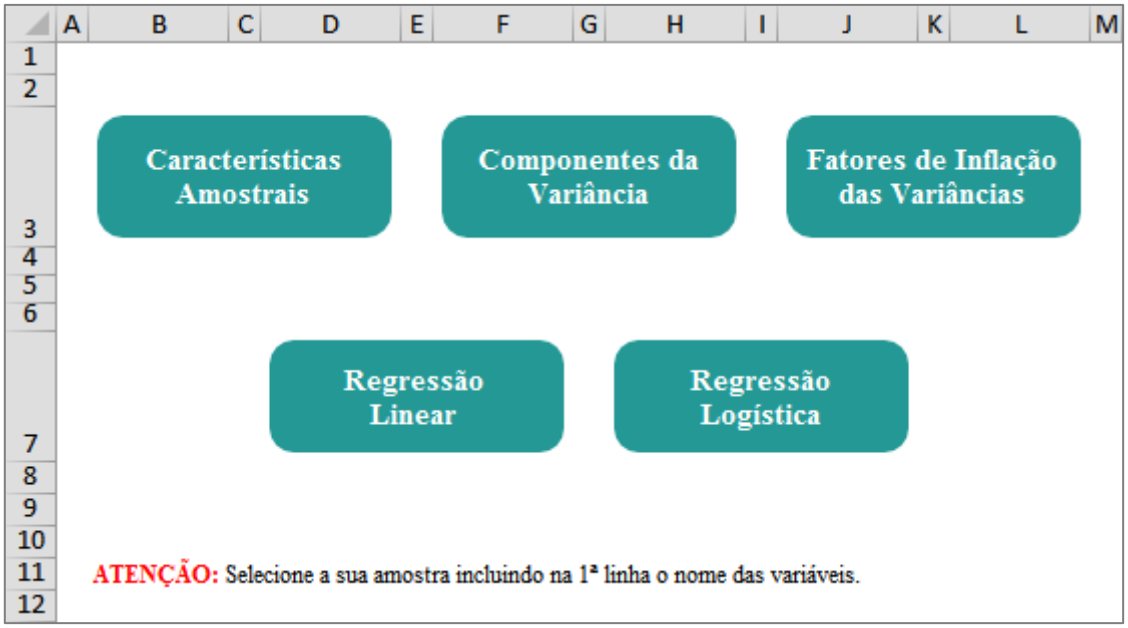

Figura 6.1 - Menu de Procedimentos

Alternativamente, o utilizador pode correr as macros num ficheiro à sua escolha, desde que o ficheiro original com as macros implementadas no âmbito deste relatório também esteja aberto, através do menu de Macros do Excel.

6.1. CARACTERÍSTICAS AMOSTRAIS

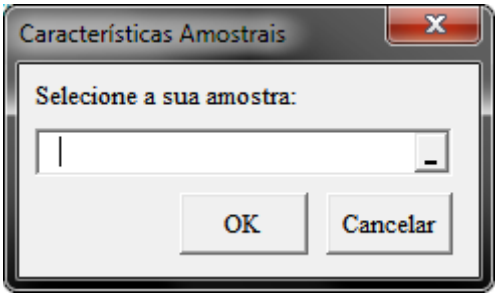

Figura 6.2 - Janela de procedimentos: Características Amostrais

Deste procedimento resulta um quadro onde se apresentam as características amostrais enunciadas no capítulo 2 deste documento.

Os resultados são apresentados numa nova folha com no nome "Out_Caracteristicas".

6.2. COMPONENTES DA VARIÂNCIA

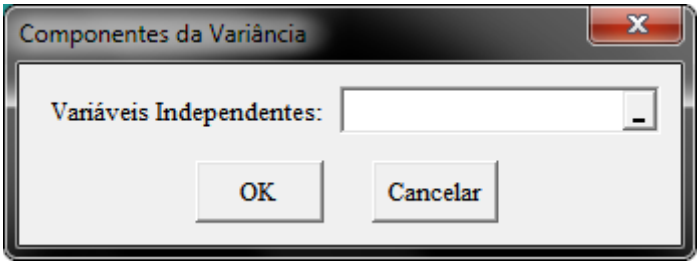

Figura 6.3 – Janela de procedimentos: Componentes da Variância

Este procedimento permite obter o resultado da metodologia descrita no capítulo 3.1 deste documento.

É relevante referir que o algoritmo implementado para determinar os valores e vetores próprios de uma matriz foi desenvolvido pelo prof. Eduardo Severino da Faculdade de Ciências da Universidade de Lisboa, tendo sido disponibilizado pela Prof. Teresa Alpuim no âmbito da cadeira de Modelos Lineares do Mestrado em Matemática Aplicada à Economia e Gestão.

Os resultados são apresentados numa nova folha com no nome "Out_Comp_Variancia".

6.3. FATORES DE INFLAÇÃO DAS VARIÂNCIAS

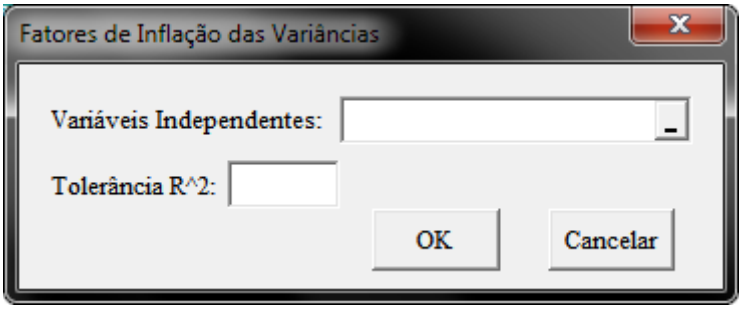

Figura 6.4 – Janela de procedimentos: Fatores de Inflação das Variâncias

Este procedimento permite obter o resultado da metodologia descrita no capítulo 3.2 deste documento.

Os resultados são apresentados numa nova folha com no nome "Out_Infl_Variancia".

6.4. REGRESSÃO LINEAR

Figura 6.5 – Janela de procedimentos: Regressão Linear

Este procedimento permite ao utilizador ajustar modelos de regressão linear a um conjunto de observações tendo a possibilidade de escolher o método de seleção de variáveis a incluir no modelo. Ao utilizador é solicitado que insira a coluna correspondente à variável dependente e a amostra que contém as variáveis independentes. Posteriormente o utilizador deverá decidir se pretende que o seu modelo seja desenvolvido com termo constante ou não. É ainda necessário escolher o nível de significância a usar para determinar a significância das variáveis. Caso o utilizador pretenda obter um modelo usando o método de seleção *stepwise* deve escolher um segundo nível de significância, correspondente ao nível a que são retiradas variáveis, tal como descrito no capítulo 5.3.

O utilizador pode ainda optar por desenvolver simultaneamente modelos através dos três métodos de seleção de variáveis implementados, tendo presente que desta forma os modelos serão todos desenvolvidos com termo constante, ou todos sem termo constante; e todos os métodos aplicam o mesmo nível de significância definido.

Para a regressão linear os estimadores de mínimos quadrados e respetivos desvios-padrão, bem como as quantidades: $\boldsymbol{R^2}$**,** $\boldsymbol{S}$**,** Estatística de Teste F, graus de liberdade, $\boldsymbol{SQ_R}$ **e** $\boldsymbol{SQ_e}$, são obtidos recorrendo à função do Excel **LINEST**. O output desta função é conforme se apresenta na Tabela 6.1.

Tabela 6.1 - Tabela de resultados da função **LINEST** do Excel

	A	B	C	D	E
1	$\hat{b}_k$	$\hat{b}_{k-1}$	…	$\hat{b}_1$	$\hat{b}_0$
2	$\hat{\sigma}(\hat{b}_k)$	$\hat{\sigma}(\hat{b}_{k-1})$	…	$\hat{\sigma}(\hat{b}_1)$	$\hat{\sigma}(\hat{b}_0)$
3	$\boldsymbol{R^2}$	$\boldsymbol{S} = \sqrt{\mathrm{SQ_e}/\mathrm{G.\,L.}}$			
4	**Est. Teste F** $= \dfrac{\mathrm{SQ_{Reg}}}{k' - 1} \Big/ \dfrac{\mathrm{SQ_e}}{n - k'}$	**G.L.** $= n - k' - 1$			
5	$\mathbf{SQ_{Reg}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \boldsymbol{SQ_e}$	$\mathbf{SQ_e} = \sum_{i=1}^{n} \varepsilon_i^2$			

Os resultados são apresentados numa nova folha com no nome "Out_Backward_Linear", "Out_Forward_Linear" ou "Out_Stepwise_Linear" dependendo do(s) método(s) de seleção de variáveis selecionado(s).

6.5. REGRESSÃO LOGÍSTICA

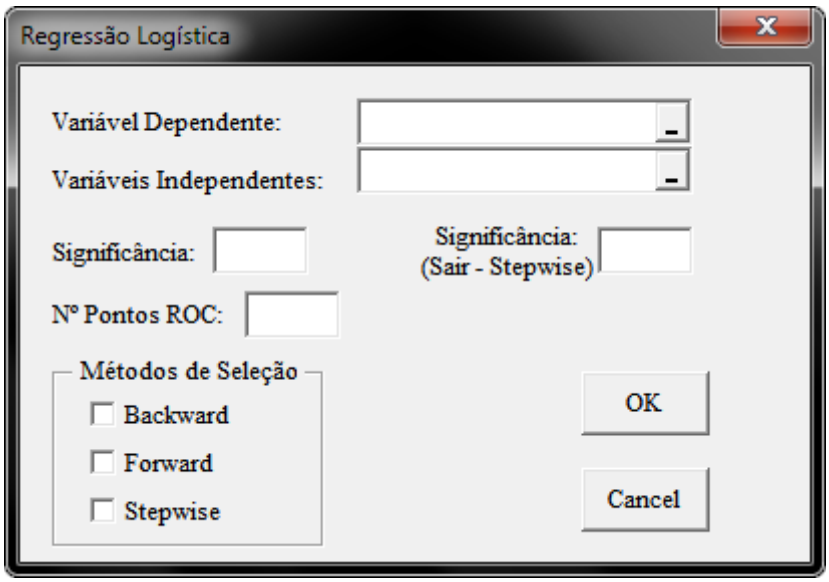

Figura 6.6 - Janela de procedimentos: Regressão Logística

Com este procedimento o utilizador pode ajustar modelos de regressão logística. À semelhança do procedimento para modelos de regressão linear, ao utilizador é solicitado que selecione a coluna correspondente à variável dependente e a amostra que contém as variáveis independentes. Caso o utilizador pretenda obter um modelo usando o método de seleção *stepwise* deve escolher um segundo nível de significância, tal como descrito no capítulo 5.3. O utilizador deve ainda escolher com quantos pontos pretende que a curva ROC seja desenhada.

Do mesmo modo que para a regressão linear, o utilizador pode ainda optar por desenvolver simultaneamente modelos através dos três métodos de seleção de variáveis implementados, tendo presente que desta forma os modelos serão todos desenvolvidos aplicando o mesmo nível de significância definido; e para todos os métodos as curvas ROC terão o mesmo número de pontos.

Para a regressão logística os estimadores de mínimos quadrados e respetivos desvios-padrão, bem como as estatísticas: $\frac{L_0}{L_1}$, $-2 \times ln\left(\frac{L_0}{L_1}\right)$ e os graus de liberdade são obtidos recorrendo a uma função implementada pela prof. Teresa Alpuim e disponibilizada no âmbito da cadeira de Modelos Lineares do Mestrado em Matemática Aplicada à Economia e Gestão, semelhante à função do Excel **LINEST** mas para a regressão logística, de onde se obtém a Tabela 6.2.

Tabela 6.2 - Tabela de resultados: Regressão Logística

	A	B	C	D	E
1	$\hat{b}_0$	$\hat{b}_1$	…	$\hat{b}_{k-1}$	$\hat{b}_k$
2	$\sigma(\hat{b}_0)$	$\sigma(\hat{b}_1)$	…	$\sigma(\hat{b}_{k-1})$	$\sigma(\hat{b}_k)$
3	"**L0/L1**"	"$-2 * ln(L0/L1)$"			
4	$\frac{L_0}{L_1}$	$-2 \times ln\left(\frac{L_0}{L_1}\right)$			
5	"**G. L.**"	k			

Os resultados serão apresentados numa nova folha com no nome "Out_Backward_Logistica", "Out_Forward_Logistica" ou "Out_Stepwise_Logistica" dependendo do(s) método(s) de seleção de variáveis selecionado(s).

6.6. NOTAS SOBRE A PROGRAMAÇÃO DE ALGUNS ALGORITMOS

6.6.1. CORRELAÇÕES

Na implementação deste processo foi necessário arredondar a dez casas decimais as entradas da matriz de correlações. Tal deve-se ao facto de que quando se está perante valores extremamente pequenos numa matriz, o VBA não consegue efetuar corretamente a inversa dessa mesma matriz, obtendo valores que não são os corretos. Efetivamente, ao ser efetuado este ajustamento o erro dos cálculos aumenta. Contudo, o erro não é suficientemente grande ao ponto de enviesar as conclusões que devem ser retiradas.

6.6.2. MÉTODO DE SELEÇÃO REGRESSIVA

Para cada iteração, são apresentados os *p-values* de todas as variáveis do modelo testado, destacando a cinza o maior *p-value* que corresponde à variável que será retirada da iteração seguinte. É também apresentado, numa coluna à direita, o valor de R^2 ou de $-2 \times ln\left(\frac{L_0}{L_1}\right)$, caso se trate de uma regressão linear ou de uma regressão logística, respetivamente.

6.6.3. MÉTODO DE SELEÇÃO PROGRESSIVA

Para cada iteração, são apresentados os *p-values* das variáveis já selecionadas incluído a variável que provoca o maior aumento de R^2 ou maior diminuição da Razão de Verosimilhanças, sobre a qual será testada a sua significância. É também apresentado, numa coluna à direita, o valor de R^2 ou de $-2 \times ln\left(\frac{L_0}{L_1}\right)$, caso se trate de uma regressão linear ou de uma regressão logística, respetivamente.

6.6.4. MÉTODO DE SELEÇÃO *STEPWISE*

Para cada iteração, são apresentados os *p-values* das variáveis já selecionadas incluído a variável que provoca o maior aumento de R^2 ou maior diminuição da Razão de Verosimilhanças, sobre a qual será testada a sua significância. É destacado a azul a variável adicionada em cada iteração, e destacadas a cinza estão as variáveis que devem ser excluídas do modelo na iteração seguinte.

É também apresentado, numa coluna à direita, o valor de R^2 ou de $-2 \times ln\left(\frac{L_0}{L_1}\right)$, caso se trate de uma regressão linear ou de uma regressão logística, respetivamente.

6.6.5. MATRIZ DE CONFUSÃO

Por defeito, o ponto de corte apresentado é o valor médio dos valores previstos. Contudo, o utilizador pode alterar este valor e desta forma obter matrizes de confusão diferentes.

6.7. LIMITAÇÕES E CUIDADOS A TER

De seguida serão enunciadas as limitações decorrentes da implementação destes programas, bem como cuidados que o utilizador deve ter na sequência dessas limitações:

- Antes de correr um procedimento, o utilizador deverá certificar-se que não existe nenhuma folha no ficheiro com o nome pré-definido para os resultados do procedimento que pretende executar.

- As variáveis devem estar dispostas em coluna.

- Quando o utilizador selecionar a(s) amostra(s) para os diferentes processos, deve sempre selecionar as células de forma a que a primeira linha contenha o nome das variáveis.

- Enquanto que para desenvolver modelos de regressão linear o utilizador pode escolher se pretende que o modelo final tenha ou não termo constante, na regressão logística tal não acontece. Como tal, na regressão logística, o modelo desenvolvido terá sempre um termo constante. O algoritmo poderá terminar com o *p-value* do termo constante superior ao nível de significância definido. Isso significa que o modelo poderá ser melhor ajustado caso não tenha termo constante.
  Contudo, apesar de ser referido como um ponto fraco da implementação dos algoritmos, é recomendado que a regressão logística inclua termo constante, uma vez que não é trivial interpretar o seu efeito, e obriga a que a probabilidade do evento quando todas as variáveis independentes são iguais a zero seja 1/2.

- Os algoritmos desenvolvidos pressupõem que as variáveis a testar são contínuas. Caso o utilizador pretenda desenvolver modelos usando variáveis categóricas, sugere-se a criação de variáveis indicatrizes (*dummy*) para esse efeito.

- Os algoritmos desenvolvidos não funcionam se o vetor da variável dependente ou a matriz de planeamento tiverem registos com informação em falta (*missings*). Deste modo, o utilizador terá que fazer um tratamento da informação *a priori* para corrigir este tipo de situações. Por exemplo, eliminar os registos ou as variáveis que tenham *missings*, ou preenchendo os *missings* com algum valor, nomeadamente com a média das observações, ou algum percentil. Este tipo de ações deve ficar à responsabilidade do utilizador uma vez que o tratamento a dar nestas situações depende do problema em questão.

- Antes de apagar uma folha com a curva ROC deve primeiro apagar-se o gráfico. Caso contrário o programa torna-se irresponsivo ao tentar editar células, não permitindo correr os programas novamente. Alternativamente, o utilizador pode, após apagar a folha, guardar o ficheiro, fechar e voltar a abrir podendo desta forma continuar com os seus exercícios. Não foi possível detetar se a origem desta limitação poderá estar nos algoritmos desenvolvidos, pelo que se conclui que poderá ser consequência de um *bug* do software Microsoft Office Excel. Adicionalmente, é apresentada uma mensagem para alertar o utilizador.

6.8. MELHORIAS

No seguimento das limitações anteriormente descritas, são agora apresentadas algumas melhorias que podem ser feitas aos algoritmos já desenvolvidos, podendo servir de base para um complemento dos programas já desenvolvidos, de forma a desenvolver um pacote de algoritmos estatísticos mais completos e eficientes.

- Uma vez que é necessário que o utilizador efetue tratamento da informação no que diz respeito aos *missings*, poderia ser útil implementar um processo que fizesse esse tratamento, tendo apenas o utilizador de escolher, de entre uma lista de opções o método preferencial para o preenchimento dos registos em falta de cada variável.

- Na presença de variáveis categóricas, poderia ser disponibilizado ao utilizador um processo que automaticamente produzisse as suas variáveis indicatrizes equivalentes correspondentes.

- A matriz de confusão apresentada para a regressão logística pode ser complementada apresentando mais alguns rácios como, por exemplo, os que se apresentam na Tabela 6.3.

Tabela 6.3 - Matriz de Confusão Complementada

		Observados			
	Nº. Obs. (n)	**Positivos**	**Negativos**	**Precisão** $= \frac{VP + VN}{n}$	
Previstos	**Positivos**	VP	FP (Erro de tipo I)	$\frac{VP}{VP + FP}$	$\frac{FP}{VP + FP}$
	Negativos	FN (Erro de tipo II)	VN	$\frac{FN}{FN + VN}$	$\frac{VN}{FN + VN}$
	Prevalência $= \frac{VP + FN}{n}$	$\mathbf{S} = \frac{VP}{VP + FN}$	$\mathbf{1 - E} = \frac{FP}{FP + VN}$	$\mathbf{RVP} = \frac{S}{1 - E}$	$\mathbf{DOR} = \frac{RVP}{RVN}$
		$\mathbf{1 - S} = \frac{FN}{VP + FN}$	$\mathbf{E} = \frac{VN}{FP + VN}$	$\mathbf{RVN} = \frac{1 - S}{E}$	

onde:

RVP: Rácio de Verosimilhança Positiva

RVN: Rácio de Verosimilhança Negativa

DOR: *Diagnostic Odds Ratio*

7. EXEMPLOS

Para efeitos de demonstração da aplicação dos algoritmos desenvolvidos, apresentam-se dois exemplos meramente ilustrativos para ajustamento de modelos de regressão linear e de regressão logística. Nestes exemplos, as amostras serão analisadas usando apenas os procedimentos implementados, e as regressões serão também ajustadas apenas com os algoritmos implementados no âmbito deste relatório.

7.1. REGRESSÃO LINEAR

Para a regressão linear, desenvolveu-se um modelo para prever a qualidade do vinho verde.

Os dados utilizados foram obtidos na internet e a amostra foi recolhida por investigadores portugueses da Universidade do Minho em 2009.

A amostra contém variáveis que são resultado de testes objetivos e a variável dependente é baseada em dados sensoriais (mediana de um conjunto de, pelo menos, 3 avaliações feitas por enólogos experientes. Cada especialista avaliou a qualidade do vinho com uma pontuação entre 0 (muito mau) e 10 (muito bom).

Por uma questão de privacidade e logística, não foram disponibilizadas variáveis referentes, por exemplo, ao tipo de uva, marca de vinho ou preço de venda.

Neste exemplo, irá ser ajustado um modelo de regressão linear à qualidade do vinho verde, tendo como ponto de partida 11 variáveis independentes contínuas com 4 898 registos.

7.1.1. CARACTERÍSTICAS AMOSTRAIS

Uma primeira análise que convém efetuar, independentemente dos modelos de regressão que posteriormente serão utilizados, é a análise das principais características amostrais. Na Figura 7.1 mostram-se as essas características para a amostra usada para exemplificar o ajustamento de uma regressão linear.

Características Amostrais

	Registos	Observações	% Missings	Média	Variância	Desvio-Padrão	Mediana	Mínimo	Máximo
fixed acidity	4 898	4 898	0,00%	6,85	0,71	0,84	6,80	3,80	14,20
volatile acidity	4 898	4 898	0,00%	0,28	0,01	0,10	0,26	0,08	1,10
citric acid	4 898	4 898	0,00%	0,33	0,01	0,12	0,32	0,00	1,66
residual sugar	4 898	4 898	0,00%	6,39	25,73	5,07	5,20	0,60	65,80
chlorides	4 898	4 898	0,00%	0,05	0,00	0,02	0,04	0,01	0,35
free sulfur dioxide	4 898	4 898	0,00%	35,31	289,24	17,01	34,00	2,00	289,00
total sulfur dioxide	4 898	4 898	0,00%	138,36	1 806,09	42,50	134,00	9,00	440,00
density	4 898	4 898	0,00%	0,99	0,00	0,00	0,99	0,99	1,04
pH	4 898	4 898	0,00%	3,19	0,02	0,15	3,18	2,72	3,82
sulphates	4 898	4 898	0,00%	0,49	0,01	0,11	0,47	0,22	1,08
alcohol	4 898	4 898	0,00%	10,51	1,51	1,23	10,40	8,00	14,20
quality	4 898	4 898	0,00%	5,88	0,78	0,89	6,00	3,00	9,00

Figura 7.1 – Características Amostrais (Ex. Regressão Linear)

Como se pode ver, a variável **density** tem uma amplitude bastante reduzida, variando as suas observações entre 0,99 e 1,04. Desta forma, e como a variável apresenta uma variância quase nula, será retirada da amostra.

7.1.2. COMPONENTES DA VARIÂNCIA

De seguida apresenta-se o quadro das componentes da variância com o objetivo de identificar possíveis relações lineares entre duas ou mais variáveis independentes a fim de reduzir os efeitos da multicolinearidade.

Componentes da Variância

Valores Próprios	Números Condição	Proporções da variância do coeficiente de regressão de										
		Const	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	pH	sulphates	alcohol
9,98	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	124,88	0,97	0,33	0,00	0,00	0,06	0,04	0,01	0,00	0,86	0,00	0,07
0,01	28,07	0,00	0,62	0,02	0,06	0,02	0,01	0,06	0,08	0,01	0,03	0,10
0,13	8,79	0,00	0,00	0,27	0,42	0,02	0,00	0,05	0,01	0,00	0,00	0,00
0,37	5,16	0,00	0,00	0,00	0,00	0,57	0,00	0,01	0,00	0,00	0,00	0,00
0,17	7,69	0,00	0,00	0,00	0,00	0,00	0,81	0,01	0,00	0,00	0,00	0,00
0,18	7,53	0,00	0,00	0,11	0,00	0,17	0,00	0,37	0,01	0,00	0,00	0,00
0,04	15,14	0,00	0,01	0,01	0,02	0,03	0,02	0,18	0,31	0,00	0,37	0,02
0,00	45,03	0,02	0,02	0,01	0,01	0,05	0,10	0,00	0,11	0,13	0,02	0,80
0,07	11,72	0,00	0,00	0,44	0,44	0,04	0,01	0,03	0,00	0,00	0,14	0,00
0,04	16,81	0,00	0,02	0,13	0,04	0,03	0,01	0,27	0,47	0,00	0,43	0,00

Figura 7.2 – Componentes da Variância (Ex. Regressão Linear)

Existem 2 números condição superiores a 30. O primeiro (124,88) mostra que as observações da variável **pH** estão próximas de uma constante, uma vez que as proporções das componentes da variância dos coeficientes de regressão das referidas variáveis são ambas próximas da unidade. Como a amplitude da variável **pH** é baixa, optou-se por retirar a variável do modelo.

Componentes da Variância

Valores Próprios	Números Condição	Proporções da variância do coeficiente de regressão de									
		Const	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	sulphates	alcohol
9,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,00	58,97	0,98	0,25	0,00	0,01	0,08	0,12	0,00	0,06	0,03	0,73
0,01	27,75	0,01	0,68	0,03	0,03	0,04	0,01	0,07	0,11	0,03	0,23
0,13	8,35	0,00	0,00	0,27	0,42	0,02	0,00	0,05	0,01	0,00	0,00
0,37	4,96	0,00	0,00	0,01	0,00	0,60	0,00	0,01	0,00	0,00	0,00
0,17	7,36	0,00	0,00	0,02	0,00	0,00	0,81	0,00	0,00	0,01	0,00
0,18	7,17	0,00	0,00	0,11	0,00	0,16	0,01	0,38	0,01	0,00	0,00
0,04	14,63	0,00	0,01	0,01	0,03	0,04	0,02	0,24	0,43	0,22	0,03
0,07	11,38	0,00	0,00	0,40	0,43	0,03	0,01	0,03	0,00	0,22	0,00
0,03	16,13	0,00	0,05	0,15	0,07	0,02	0,01	0,21	0,37	0,49	0,01

Figura 7.3 – Componentes da Variância (Ex. Regressão Linear)

Retirando a variável **pH** obtém-se apenas um número condição superior a 30 causado pelo efeito da variável **alcohol** na regressão. Apesar de se poder concluir que esta variável tem observações próximas de uma constante, a sua variância não é muito reduzida (1,51), pelo que não faz sentido retirar a variável com base nesta análise.

7.1.3. FATORES DE INFLAÇÃO DAS VARIÂNCIAS

Na Figura 7.4 pode ver-se o resultado do método dos fatores de inflação das variâncias.

Fatores de Inflação das Variâncias

Tolerância R^2: 0,500
Tolerância VIF: 2,000

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	sulphates	alcohol
fixed acidity	1,123	-0,022	-0,333	-0,044	0,056	0,126	0,032	0,145
volatile acidity	-0,022	1,071	0,176	-0,179	-0,152	0,112	0,015	-0,169
citric acid	-0,333	0,176	1,154	-0,091	-0,148	-0,082	-0,068	-0,081
residual sugar	-0,044	-0,179	-0,091	1,369	0,146	-0,289	0,060	0,598
chlorides	0,056	-0,152	-0,148	0,146	1,195	-0,040	0,000	0,492
free sulfur dioxide	0,126	0,112	-0,082	-0,289	-0,040	1,155	-0,062	0,145
sulphates	0,032	0,015	-0,068	0,060	0,000	-0,062	1,011	0,027
alcohol	0,145	-0,169	-0,081	0,598	0,492	0,145	0,027	1,506

Figura 7.4 – Fatores de Inflação das Variâncias (Ex. Regressão Linear)

Com uma tolerância para R^2 de 0,5, ou seja, para um VIF superior a 2, o resultado da aplicação deste método conduz à conclusão de que a variável **total sulfur dioxide** estaria a contribuir para um efeito de multicolinearidade, e deste modo deve ser retirada do ajustamento.

Uma vez que já foram retiradas variáveis causadoras de multicolinearidade, pode então proceder-se ao ajustamento de uma regressão linear. Optou-se por usar um nível de significância de 0,05 para os métodos de seleção de variáveis *Backward* e *Forward*, e níveis de 0,05 e 0,1 para o método de seleção de variáveis Stepwise. Estes valores foram escolhidos arbitrariamente, contudo é bastante frequente o seu uso na *literatura*, em situações semelhantes.

7.1.4. REGRESSÃO LINEAR – *BACKWARD*

Na regressão linear, o utilizador pode escolher se pretende um modelo com ou sem termo constante. Uma forma de ver qual dos métodos se ajustará melhor será comparar a eficácia do ajustamento em ambas as situações, mantendo constante o nível de significância e o método de seleção de variáveis usados.

Para este exemplo em particular, ajustaram-se dois modelos (com e sem termo constante) com nível de significância de 5% e usando o método de seleção regressiva. Apresentam-se na Figura 7.5 e na Figura 7.6 os resultados de ambos os ajustamentos.

Regressão Linear: Método de Seleção Regressiva (Backward Selection)

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	sulphates	alcohol	const
0,000	0,000	0,613	0,000	0,047	0,000	0,000	0,000	0,000
0,000	0,000		0,000	0,039	0,000	0,000	0,000	0,000

R^2
0,272
0,271

Coeficientes	fixed acidity	volatile acidity	residual sugar	chlorides	free sulfur dioxide	sulphates	alcohol	const
EMQ	-0,067	-2,014	0,024	-1,110	0,004	0,414	0,370	2,577
σ(b)	0,013	0,110	0,002	0,536	0,001	0,095	0,011	0,176
Estat. Teste	-5,174	-18,377	9,644	-2,070	5,334	4,354	34,361	14,644
P-Value	0,000	0,000	0,000	0,039	0,000	0,000	0,000	0,000

α	0,050

R^2	0,271
Teste-F	260,306
SQ_{Reg}	1 042,712

S	0,756
G.L.	4 890
SQ_e	2 798,278

Figura 7.5 – Seleção Regressiva com termo constante (Ex. Regressão Linear)

Regressão Linear: Método de Seleção Regressiva (Backward Selection)

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	sulphates	alcohol	const	R^2
0,000	0,000	0,319	0,000	0,001	0,000	0,000	0,000		0,983
0,000	0,000		0,000	0,001	0,000	0,000	0,000		0,983

Coeficientes	fixed acidity	volatile acidity	residual sugar	chlorides	free sulfur dioxide	sulphates	alcohol	const
EMQ	0,044	-1,853	0,035	1,720	0,006	0,805	0,492	0,000
σ(b)	0,011	0,111	0,002	0,511	0,001	0,093	0,007	
Estat. Teste	4,110	-16,634	14,352	3,364	8,728	8,644	71,253	
P-Value	0,000	0,000	0,000	0,001	0,000	0,000	0,000	

α	0,050

R^2	0,983
Teste-F	40 699,308
SQ_{Reg}	170 144,998

S	0,773
G.L.	4 891
SQ_e	2 921,002

Figura 7.6 – Seleção Regressiva sem termo constante (Ex. Regressão Linear)

Para o modelo sem termo constante, o valor de R^2 calcula-se de forma diferente do caso do modelo com constante. Por essa razão, neste caso, os dois indicadores não são comparáveis. No entanto, pode ver-se que o valor de SQe para o modelo com constante é inferior ao SQe do modelo sem constante e, portanto, considera-se que o ajustamento com termo constante é mais fiel às observações.

7.1.5. REGRESSÃO LINEAR – *FORWARD*

Pelas conclusões obtidas no ajustamento pelo método de seleção regressiva, apresenta-se de seguida apenas o resultado obtido do ajustamento pelo método de seleção progressiva com termo constante, e usando o mesmo nível de significância. Desta forma, os métodos são comparáveis permitindo determinar qual seria o melhor método para a amostra usada.

Regressão Linear: Método de Seleção Progressiva (Forward Selection)

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	sulphates	alcohol	const	R^2
							0,000	0,000	0,190
	0,000						0,000	0,000	0,240
	0,000		0,000				0,000	0,000	0,259
	0,000		0,000		0,000		0,000	0,000	0,264
0,000	0,000		0,000		0,000		0,000	0,000	0,268
0,000	0,000		0,000		0,000	0,000	0,000	0,000	0,271
0,000	0,000		0,000	0,039	0,000	0,000	0,000	0,000	0,271
0,000	0,000	0,613	0,000	0,047	0,000	0,000	0,000	0,000	0,272

Coeficientes	alcohol	volatile acidity	residual sugar	free sulfur dioxide	fixed acidity	sulphates	chlorides	const
EMQ	0,370	-2,014	0,024	0,004	-0,067	0,414	-1,110	2,577
σ(b)	0,011	0,110	0,002	0,001	0,013	0,095	0,536	0,176
Estat. Teste	34,361	-18,377	9,644	5,334	-5,174	4,354	-2,070	14,644
P-Value	0,000	0,000	0,000	0,000	0,000	0,000	0,039	0,000

α	0,050

R^2	0,271
Teste-F	260,306
SQ_{Reg}	1 042,712

S	0,756
G.L.	4 890
SQ_e	2 798,278

Figura 7.7 – Seleção Progressiva com termo constante (Ex. Regressão Linear)

7.1.6. REGRESSÃO LINEAR - *STEPWISE*

Pelas conclusões obtidas no ajustamento pelo método de seleção regressiva, apresenta-se na Figura 7.8 apenas o resultado obtido do ajustamento pelo método de seleção *stepwise* com termo constante, usando o mesmo nível de significância para avaliar as variáveis a entrar no ajustamento (5%), e o seu dobro para avaliar se as variáveis deixam de ser significativas à medida que se vão adicionando variáveis (1%).

Regressão Linear: Método de Seleção Progressiva e Regressiva (Stepwise Selection)

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	sulphates	alcohol	const	R^2
							0,000	0,000	0,190
	0,000						0,000	0,000	0,240
	0,000		0,000				0,000	0,000	0,259
	0,000		0,000		0,000		0,000	0,000	0,264
0,000	0,000		0,000		0,000		0,000	0,000	0,268
0,000	0,000		0,000		0,000	0,000	0,000	0,000	0,271
0,000	0,000		0,000	0,039	0,000	0,000	0,000	0,000	0,271
0,000	0,000	0,613	0,000	0,047	0,000	0,000	0,000	0,000	0,272

Coeficientes	alcohol	volatile acidity	residual sugar	free sulfur dioxide	fixed acidity	sulphates	chlorides	const
EMQ	0,370	-2,014	0,024	0,004	-0,067	0,414	-1,110	2,577
σ(b)	0,011	0,110	0,002	0,001	0,013	0,095	0,536	0,176
Estat. Teste	34,361	-18,377	9,644	5,334	-5,174	4,354	-2,070	14,644
P-Value	0,000	0,000	0,000	0,000	0,000	0,000	0,039	0,000

α_{in}	0,050
R^2	0,271
Teste-F	260,306
SQ_{Reg}	1 042,712

α_{out}	0,100
S	0,756
G.L.	4 890
SQ_e	2 798,278

Figura 7.8 – Seleção *Stepwise* com termo constante (Ex. Regressão Logistica)

Note-se que, para um ajustamento com termo constante, considerando os mesmos níveis de significância, se obtém exatamente o mesmo ajustamento independentemente do método de seleção das variáveis. Esta situação acontece porque as variáveis explicativas estão próximas da independência entre si. Desta forma, qualquer método iterativo de seleção de variáveis, conduzirá ao mesmo modelo.

7.2. REGRESSÃO LOGÍSTICA

Para a regressão logística, desenvolveu-se um modelo para prever o sucesso de campanhas de telemarketing para venda de um produto financeiro (depósitos a prazo) num banco português.

A amostra contém várias variáveis, contudo nem todas são contínuas. Para essas, foi necessário desagregar as variáveis categóricas em variáveis indicatrizes. A identificação da adesão do cliente ao produto exprime-se numa variável binária, assumindo sempre os valores "1" ou "0", consoante o cliente subscreveu ou não o produto financeiro.

Neste exemplo, irá ser ajustado um modelo de regressão logística ao sucesso de campanhas de telemarketing, tendo como ponto de partida 20 variáveis independentes, das quais 10 são categóricas, e 4 119 registos.

Na Tabela 7.1 apresentam-se as descrições das variáveis desta amostra.

Tabela 7.1 – Tabela descritiva das variáveis da amostra usada para ajustar uma regressão logística

Variável	Descrição	Categorias
y	Variável dependente: Indicador de cliente que subscreveu a um depósito a prazo	N.A.
age	Idade	N.A.
job	Tipo de profissão	1: admin 2: blue-collar 3: entrepreneur 4: housemaid 5: management 6: retired 7: self-employed 8: services 9: student 10: technician 11: unemployed 12: unknown
marital	Estado Civil	1: divorced (or widowed) 2: married 3: single 4: unknown
education	Nível de formação académica	1: basic.4y 2: basic.6y 3: basic.9y 4: high.school 5: illiterate 6: professional.course 7: university.degree 8: unknown
default	Indicador de crédito em incumprimento	0: no 1: yes 2: unknown
housing	Indicador de crédito à habitação	0: no 1: yes 2: unknown

Variável	Descrição	Categorias
loan	Indicador de crédito ao consumo	0: no 1: yes 2: unknown
contact	Tipo de contacto	0: celular 1: telephone
month	Mês do contacto	1: jan … 12: dec
day_of_week	Dia da semana do contacto	1: mon … 5: fri
duration[5]	Duração do contacto (sem segundos)	N.A.
campaign	Número de contactos feitos para o cliente no âmbito da campanha	N.A.
pdays	Número de dias que passaram desde o último contacto ao cliente no âmbito de outra campanha	N.A.
previous	Número de contactos feitos ao cliente no antes da campanha	N.A.
poutcome	Resultado da última campanha	0: failure 1: success 2: nonexistent
emp.var.rate	Variação da taxa de empregabilidade (resultados trimestrais)	N.A.
cons.price.idx	Índice de Preços no Consumidor (resultados mensais)	N.A.
Cons.conf.idx	Índice de Confiança do Consumidor (resultados mensais)	N.A.
euribor3m	Taxa Euribor a 3 meses (resultados diários)	N.A.
nr.employed	Média de número de cidadãos empregados (resultados trimestrais)	N.A.

Os autores deixam a nota que existem categorias com informação em falta (*missing*) que foram preenchidas com a categoria "unknown". Para efeitos deste exercício, estes valores serão tratados como uma nova categoria. Também as variáveis cateróricas foram desdobradas nas respetivas variáveis indicatrizes correspondentes.

[5] De acordo com as indicações dos autores da base de dados usada para este exercício, esta variável não será incluída no modelo: "this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. "

7.2.1. CARACTERÍSTICAS AMOSTRAIS

Numa análise preliminar, devem ser analisadas as características amostrais das variáveis que serão utilizadas para descrever a variável resposta, através de um modelo de regressão logística.

Características Amostrais

	Registos	Observações	% Missings	Média	Variância	Desvio-Padrão	Mediana	Mínimo	Máximo
y	4 119	4 119	0,00%	0,11	0,10	0,31	0,00	0,00	1,00
age	4 119	4 119	0,00%	40,11	106,37	10,31	38,00	18,00	88,00
job	4 119	4 119	0,00%	4,82	13,01	3,61	4,00	1,00	12,00
marital	4 119	4 119	0,00%	2,18	0,37	0,61	2,00	1,00	4,00
education	4 119	4 119	0,00%	4,78	4,62	2,15	4,00	1,00	8,00
default	4 119	4 119	0,00%	0,39	0,63	0,79	0,00	0,00	2,00
housing	4 119	4 119	0,00%	0,58	0,29	0,54	1,00	0,00	2,00
loan	4 119	4 119	0,00%	0,21	0,22	0,47	0,00	0,00	2,00
contact	4 119	4 119	0,00%	0,36	0,23	0,48	0,00	0,00	1,00
month	4 119	4 119	0,00%	6,69	4,26	2,06	6,00	3,00	12,00
day_of_week	4 119	4 119	0,00%	2,96	1,99	1,41	3,00	1,00	5,00
campaign	4 119	4 119	0,00%	2,54	6,60	2,57	2,00	1,00	35,00
pdays	4 119	4 119	0,00%	960,42	36 834,36	191,92	999,00	0,00	999,00
previous	4 119	4 119	0,00%	0,19	0,29	0,54	0,00	0,00	6,00
poutcome	4 119	4 119	0,00%	1,75	0,41	0,64	2,00	0,00	2,00
emp.var.rate	4 119	4 119	0,00%	0,08	2,44	1,56	1,10	-3,40	1,40
cons.price.idx	4 119	4 119	0,00%	93,58	0,34	0,58	93,75	92,20	94,77
cons.conf.idx	4 119	4 119	0,00%	-40,50	21,11	4,59	-41,80	-50,80	-26,90
euribor3m	4 119	4 119	0,00%	3,62	3,01	1,73	4,86	0,64	5,05
nr.employed	4 119	4 119	0,00%	5 166,48	5 426,96	73,67	5 191,00	4 963,60	5 228,10

Figura 7.9 – Características Amostrais (Ex. Reg. Logística)

Para este conjunto de dados, destacam-se a azul e cinza as variáveis categóricas, para as quais esta análise não trás qualquer informação adicional. Para as restantes variáveis, é fácil identificar que, à primeira vista, nenhuma variável tem uma variância suficientemente pequena para ser excluída da regressão nesta fase.

Uma vez que a amostra contém variáveis categóricas com diversas categorias, o que implica a "criação" de muitas variáveis adicionais (as variáveis indicatrizes), por limitação da capacidade de processamento do computador usado para a produção deste relatório, e dado que este exemplo é meramente indicativo das funcionalidades dos algoritmos implementados, decidiu-se não usar as variáveis categóricas com mais do que três categorias, indicadas a cinza.

7.2.2. COMPONENTES DA VARIÂNCIA

De seguida procede-se à análise das componentes da variância com o objetivo de identificar possíveis relações lineares entre duas ou mais variáveis independentes a fim de reduzir os efeitos da multicolinearidade.

Componentes da Variância

Valores Próprios	Números Condição	Proporções da variância do coeficiente de regressão de																		
		Const	age	d0	d1	h0	h1	l0	l1	contact	campaign	pdays	previous	pout0	pout1	emp.var. rate	cons. price.idx	cons. conf.idx	euribor 3m	nr. employed
0,00	67,76	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,71	0,16	0,05	0,54	0,08	0,00	0,07	0,07	0,00
10,66	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,04	17,29	0,00	0,16	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,01	0,00	0,00	0,02	0,00	0,00
2,23	2,19	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,02	0,02	0,01	0,00	0,00	0,00	0,00	0,00
0,00	40 846 826,32	0,00	0,00	0,00	0,00	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1,08	3,15	0,00	0,00	0,00	0,09	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00
0,18	7,79	0,00	0,02	0,88	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,99	3,29	0,00	0,00	0,00	0,73	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,00
0,12	9,60	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,70	0,77	0,17	0,00	0,00	0,00	0,00	0,00
0,44	4,90	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,18	0,57	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00
0,06	13,80	0,00	0,77	0,06	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,00	0,00
0,93	3,38	0,00	0,00	0,00	0,03	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,49	4,66	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,32	0,41	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
1,04	3,20	0,00	0,00	0,00	0,13	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,05	0,05	0,00	0,00	0,00	0,00	0,00
0,73	3,82	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,00	0,05	0,00	0,00	0,05	0,06	0,00	0,01	0,00	0,00	0,00	0,00
0,00	3 226,30	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,19	0,00	0,00	0,00	0,00	0,00	0,21	0,89	0,46	0,25	0,77
0,00	49,05	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,28	0,00	0,00	0,20	0,28	0,00	0,24	0,20	0,00
0,01	28,10	0,00	0,02	0,01	0,00	0,00	0,00	0,00	0,00	0,02	0,00	0,00	0,01	0,00	0,00	0,15	0,00	0,16	0,09	0,00
0,00	819,08	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,20	0,00	0,00	0,04	0,02	0,00	0,24	0,11	0,05	0,40	0,23

Figura 7.10 – Componentes da Variância (Ex. Regressão Logística)

Como se pode ver na figura acima, existem dois números condição extremamente elevados. O primeiro (40 846 826, 32) indica a forte relação entre as variáveis **h0, h1, l0** e **l1** (variáveis indicatrizes das variáveis **housing** e **loan**, respetivamente). Observando o comportamento de ambas as variáveis optou-se por remover a variável **loan** da regressão. O segundo número condição identificado (3 226,30) mostra a evidência de que as variáveis **cons.price.idx** e **nr.employed** estão muito relacionadas entre si, uma vez que as proporções da variância dos seus coeficientes de regressão estão muito próximos da unidade. Teoricamente, esta é uma conclusão que faz sentido, uma vez que o Índice de Preços no Consumidor tende a variar no mesmo sentido do que o número de trabalhadores empregados. Este resultado está diretamente relacionado com o princípio básico da economia que estabelece uma relação entre a inflação e o desemprego. Desta forma, e como a variância da variável **cons.price.idx** é menor que a variância da variável **nr.employed**, optou-se por retirar a primeira variável da regressão.

Componentes da Variância

Valores Próprios	Números Condição	Proporções da variância do coeficiente de regressão de															
		Const	age	d0	d1	h0	h1	contact	campaign	pdays	previous	pout0	pout1	emp. var.rate	cons. conf.idx	euribor 3m	nr. employed
0,00	44,94	0,00	0,01	0,00	0,00	0,00	0,00	0,01	0,00	0,39	0,00	0,00	0,28	0,37	0,31	0,19	0,00
8,70	1,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
0,17	7,12	0,00	0,03	0,86	0,00	0,00	0,00	0,02	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2,23	1,98	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,02	0,02	0,01	0,01	0,00	0,00	0,00
0,05	13,11	0,00	0,89	0,10	0,00	0,04	0,03	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,01	0,00	0,00
1,05	2,88	0,00	0,00	0,00	0,28	0,00	0,00	0,00	0,00	0,00	0,00	0,05	0,04	0,00	0,00	0,00	0,00
0,49	4,21	0,00	0,00	0,00	0,00	0,00	0,00	0,48	0,38	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00
0,44	4,45	0,00	0,00	0,01	0,00	0,00	0,00	0,25	0,60	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00
0,00	68,61	0,00	0,02	0,01	0,00	0,02	0,02	0,00	0,00	0,60	0,14	0,04	0,46	0,19	0,16	0,11	0,00
0,97	2,99	0,00	0,00	0,00	0,58	0,01	0,01	0,00	0,00	0,00	0,00	0,01	0,01	0,00	0,00	0,00	0,00
0,12	8,68	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,70	0,77	0,17	0,00	0,00	0,00	0,00
1,01	2,93	0,00	0,00	0,00	0,13	0,02	0,01	0,01	0,00	0,00	0,00	0,01	0,02	0,00	0,00	0,00	0,00
0,72	3,48	0,00	0,00	0,01	0,01	0,00	0,00	0,07	0,00	0,00	0,05	0,06	0,00	0,02	0,00	0,00	0,00
0,02	20,15	0,00	0,01	0,00	0,00	0,89	0,90	0,00	0,00	0,01	0,00	0,00	0,01	0,00	0,03	0,00	0,00
0,01	25,46	0,00	0,03	0,01	0,00	0,01	0,01	0,02	0,00	0,00	0,01	0,00	0,00	0,25	0,22	0,10	0,00
0,00	1 193,65	1,00	0,00	0,00	0,00	0,00	0,00	0,12	0,00	0,00	0,04	0,02	0,00	0,14	0,26	0,60	1,00

Figura 7.11 – Componentes da Variância (Ex. Regressão Logística)

Após a remoção das variáveis **loan** e **cons.price.idx** permanece um número condição demasiado elevado. Apesar de se pretender concluir que a variável **nr.employed** terá observações próximas de uma constante, a sua variância não é muito reduzida, pelo que não faz sentido retirar a variável com base nesta análise. Uma análise mais detalhada leva à conclusão que este efeito é causado pelo facto de a variável apenas apresentar onze valores distintos (em 4 119 registos).

7.2.3. FATORES DE INFLAÇÃO DAS VARIÂNCIAS

Na Figura 7.12 pode ver-se o resultado do método dos fatores de inflação das variâncias.

Fatores de Inflação das Variâncias

Tolerância R^2: 0,600
Tolerância VIF: 2,500

	age	d0	d1	h0	contact	campaign	pout0	pout1	cons.conf.idx	nr.employed
age	1,045	0,184	0,020	-0,009	0,020	0,005	-0,011	-0,026	-0,110	0,067
d0	0,184	1,077	0,036	-0,006	0,131	-0,011	0,017	-0,027	-0,039	0,157
d1	0,020	0,036	1,004	-0,018	0,012	0,003	-0,055	-0,013	-0,004	-0,027
h0	-0,009	-0,006	-0,018	1,006	-0,064	-0,011	-0,003	0,014	-0,012	-0,006
contact	0,020	0,131	0,012	-0,064	1,179	-0,052	0,149	0,089	-0,260	-0,162
campaign	0,005	-0,011	0,003	-0,011	-0,052	1,030	0,020	-0,003	0,024	-0,150
pout0	-0,011	0,017	-0,055	-0,003	0,149	0,020	1,253	0,270	0,091	0,504
pout1	-0,026	-0,027	-0,013	0,014	0,089	-0,003	0,270	1,227	-0,139	0,519
cons.conf.idx	-0,110	-0,039	-0,004	-0,012	-0,260	0,024	0,091	-0,139	1,111	-0,082
nr.employed	0,067	0,157	-0,027	-0,006	-0,162	-0,150	0,504	0,519	-0,082	1,473

Figura 7.12 - Fatores de Inflação das Variâncias (Ex. Regressão Logística)

Com uma tolerância para R^2 de 0,6, o resultado da aplicação deste método conduz à conclusão de que as variáveis **h1, pdays, previous, emp.var.rate** e **Euribor.3m** estariam a contribuir para um efeito de multicolinearidade, e deste modo devem ser retiradas do ajustamento.

Uma vez que já foram retiradas variáveis causadoras de multicolinearidade, pode então proceder-se ao ajustamento de uma regressão logística. À semelhança do que foi feito para o exemplo de regressão linear, também para este exemplo se usou um nível de significância de 0,05 para os métodos de seleção de variáveis *Backward* e *Forward*, e níveis de 0,05 e 0,1 para o método de seleção de variáveis *Stepwise*.

É relevante referir que numa primeira tentativa de aplicação do algoritmo, foi devolvida uma mensagem de erro, não permitindo fazer o exercício até ao fim. Uma análise mais aprofundada permitiu concluir que a causa do erro estava na variável indicatriz **pout1** uma vez que mais de 96% das suas observações são nulas, não permitindo que os estimadores de máxima verosimilhança fossem calculados. Uma vez retirada a variável, os algoritmos correram conforme esperado, onde, de seguida, se apresentam os resultados obtidos.

7.2.4. REGRESSÃO LOGÍSTICA – *BACKWARD*

De seguida apresenta-se o resultado do ajustamento de regressão logística usando o método de seleção regressiva, bem como a respetiva curva ROC.

Regressão Logística: Método de Seleção Regressiva (Backward Selection)

age	d0	d1	h0	contact	campaign	pout0	cons. conf.idx	nr. employed	const
0,038	0,363	0,989	0,444	0,000	0,034	0,000	0,000	0,000	0,000
0,038	0,361		0,445	0,000	0,034	0,000	0,000	0,000	0,000
0,038	0,359			0,000	0,035	0,000	0,000	0,000	0,000
0,051				0,000	0,035	0,000	0,000	0,000	0,000
				0,000	0,035	0,000	0,000	0,000	0,000

$-2*\ln(L_0/L_1)$
494,275
494,211
493,628
492,765
488,968

Coeficientes	contact	campaign	pout0	cons. conf.idx	nr. employed	const
EMQ	-0,704	-0,071	-0,715	0,039	-0,012	63,047
σ(b)	0,145	0,033	0,160	0,010	0,001	3,481
Estat. Teste	-4,847	-2,110	-4,469	4,066	-17,905	18,114
P-Value	0,000	0,035	0,000	0,000	0,000	0,000

α	0,050

L_0/L_1	0,000
$-2*\ln(L_0/L_1)$	488,968
G.L.	5

Cut-off	0,109

		Observados		
		Positivos	Negativos	Total
Previstos	Positivos	294	754	1048
	Negativos	157	2914	3071
	Total	451	3668	4119

Figura 7.13 – Seleção Regressiva (Ex. Regressão Logística)

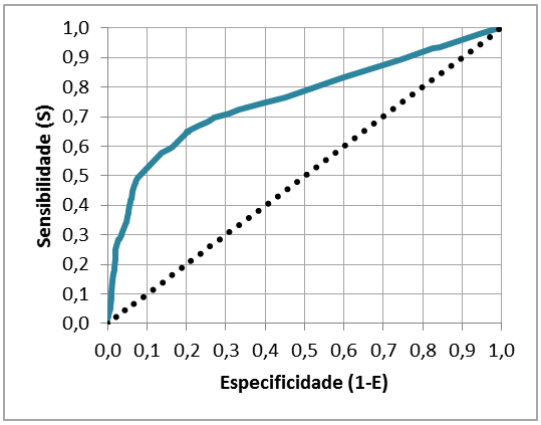

Figura 7.14 – Curva ROC do ajustamento usando a seleção regressiva (Ex. Regressão Logística)

7.2.5. REGRESSÃO LOGÍSTICA – *FORWARD*

A Figura 7.15 e a Figura 7.16 referem-se ao ajustamento de regressão logística usando o método de seleção progressiva.

Regressão Logística: Método de Seleção Progressiva (Forward Selection)

age	d0	d1	h0	contact	campaign	pout0	cons.conf.idx	nr.employed	const	-2*ln(L_0/L_1)
								0,000	0,000	429,989
						0,000		0,000	0,000	448,141
				0,000		0,000		0,000	0,000	466,758
				0,000		0,000	0,000	0,000	0,000	483,879
				0,000	0,035	0,000	0,000	0,000	0,000	488,968
0,051				0,000	0,035	0,000	0,000	0,000	0,000	492,765

Coeficientes	nr. employed	pout0	contact	cons. conf.idx	campaign	const
EMQ	-0,012	-0,715	-0,704	0,039	-0,071	63,047
σ(b)	0,001	0,160	0,145	0,010	0,033	3,481
Estat. Teste	-17,905	-4,469	-4,847	4,066	-2,110	18,114
P-Value	0,000	0,000	0,000	0,000	0,035	0,000

α	0,050

L_0/L_1	0,000
-2*ln(L_0/L_1)	488,968
G.L.	5

Cut-off	0,109

		Observados		
		Positivos	Negativos	Total
Previstos	Positivos	294	754	1048
	Negativos	157	2914	3071
	Total	451	3668	**4119**

Figura 7.15 - Seleção Progressiva (Ex. Regressão Logística)

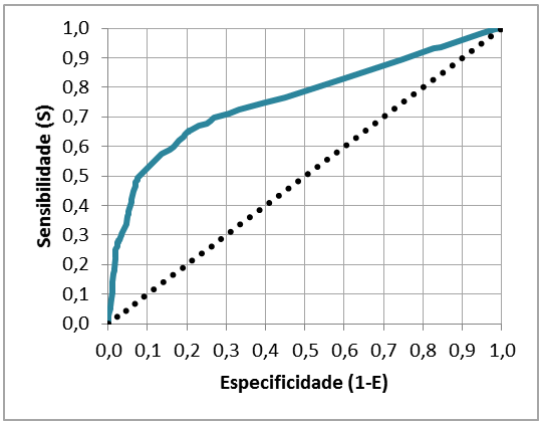

Figura 7.16 - Curva ROC do ajustamento usando a seleção progressiva (Ex. Regressão Logística)

7.2.6. REGRESSÃO LOGÍSTICA - *STEPWISE*

Regressão Logística: Método de Seleção Progressiva e Regressiva (Stepwise Selection)

age	d0	d1	h0	contact	campaign	pout0	cons. conf.idx	nr. employed	const	-2*ln(L_0/L_1)
								0,000	0,000	429,989
						0,000		0,000	0,000	448,141
				0,000		0,000		0,000	0,000	466,758
				0,000		0,000	0,000	0,000	0,000	483,879
				0,000	0,035	0,000	0,000	0,000	0,000	488,968
0,051				0,000	0,035	0,000	0,000	0,000	0,000	492,765

Coeficientes	nr. employed	pout0	contact	cons. conf.idx	campaign	const
EMQ	-0,012	-0,715	-0,704	0,039	-0,071	63,047
σ(b)	0,001	0,160	0,145	0,010	0,033	3,481
Estat. Teste	-17,905	-4,469	-4,847	4,066	-2,110	18,114
P-Value	0,000	0,000	0,000	0,000	0,035	0,000

α_{in}	0,050
α_{out}	0,100
L_0/L_1	0,000
-2*ln(L_0/L_1)	488,968
G.L.	5

Cut-off	0,109

		Observados		
		Positivos	Negativos	Total
Previstos	Positivos	294	754	1048
	Negativos	157	2914	3071
	Total	451	3668	**4119**

Figura 7.17 - Seleção *Stepwise* (Ex. Regressão Logística)

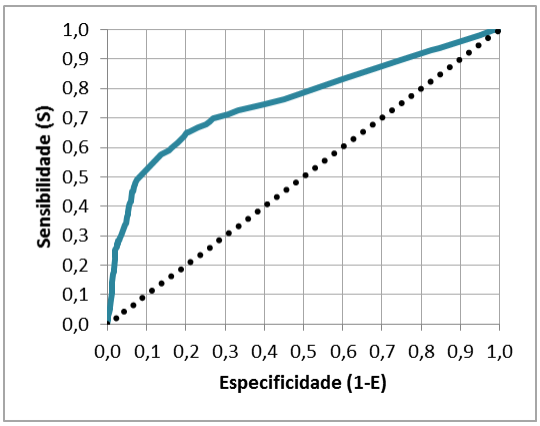

Figura 7.18 - Curva ROC do ajustamento usando a seleção *stepwise* (Ex. Regressão Logística)

7.2.7. REGRESSÃO LOGÍSTICA – *STEPWISE (VARIÁVEIS CONTÍNUAS)*

A título explicativo, e para permitir visualizar a diferença entre o método de seleção progressivo e *stepwise*, apresenta-se o resultado do ajustamento de regressão logística usando apenas as variáveis contínuas da amostra. Na figura seguinte é possível identificar uma iteração em que, após a introdução de uma variável significativa no modelo, outra deixou de ser significativa, tendo que ser retirada do ajustamento.

Regressão Logística: Método de Seleção Progressiva e Regressiva (Stepwise Selection)

age	contact	campaign	previous	emp. var.rate	cons. price.idx	cons. conf.idx	euribor 3m	nr.employed	const	-2*ln(L_0/L_1)
								0,000	0,000	429,989
	0,000							0,000	0,000	443,656
	0,000					0,000		0,000	0,000	462,570
	0,000				0,001	0,000		0,000	0,001	474,433
	0,000			0,000	0,000	0,000		0,579	0,000	505,416
	0,000	0,033		0,000	0,000	0,000			0,000	510,371
0,046	0,000	0,034		0,000	0,000	0,000			0,000	514,344
0,051	0,000	0,040	0,076	0,000	0,000	0,000			0,000	517,523

Coeficientes	contact	cons. conf.idx	cons. price.idx	emp. var.rate	campaign	age	const
EMQ	-1,104	0,071	1,574	-0,824	-0,071	0,009	-146,657
σ(b)	0,161	0,010	0,122	0,046	0,034	0,005	11,310
Estat. Teste	-6,865	6,761	12,895	-17,817	-2,120	1,996	-12,968
P-Value	0,000	0,000	0,000	0,000	0,034	0,046	0,000

α_{in}	0,050
α_{out}	0,100
L_0/L_1	0,000
-2*ln(L_0/L_1)	514,344
G.L.	6

Cut-off	0,109

		Observados		
		Positivos	Negativos	Total
Previstos	Positivos	295	821	1116
	Negativos	156	2847	3003
	Total	451	3668	**4119**

Figura 7.19 - Seleção *Stepwise* usando apenas variáveis contínuas (Ex. Regressão Logística)

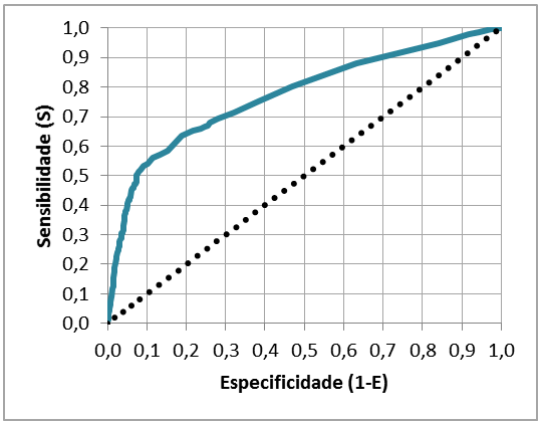

Figura 7.20 - Curva ROC do ajustamento usando a seleção *stepwise* e apenas variáveis contínuas (Ex. Regressão Logística)

Isto acontece, porque ao introduzir variáveis que, como se tinha visto, provocam multicolinearidade, o método *stepwise* tem mais dificuldade em encontrar o melhor conjunto de variáveis explicativas. Quando o conjunto de variáveis inicial não sofre de qualquer problema de multicolinearidade, o método *stepwise* tende a comportar-se do mesmo modo que o método *forward*.

3ª PARTE - BIBLIOGRAFIA E ANEXOS

8. REFERÊNCIAS BIBLIOGRÁFICAS

[1] Alpuim, T. (2016). Apontamentos da disciplina de Modelos Lineares.

[2] Alpuim, T. (2010). Apontamentos da disciplina de Estatística.

[3] Brison, O. (2010). Apontamentos da disciplina de Álgebra Linear e Geometria Analítica.

[4] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. e Reis, J.(2009). Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

[5] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. e Reis, J. (2009). Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal. Acedido a 22 de Dezembro de 2016 em: http://archive.ics.uci.edu/ml/datasets/Wine+Quality.

[6] Fahrmeir, L. e Kaupmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, The Anuals of Statistics, Vol. 18, No. 1, 342-368.

[7] Frees, E. W., Derrig, R. A. e Meyers, G. (2014). Predictive Modeling Applications in Actuarial Science: Volume 1, Predictive Modeling Techniques. 1ª edição, Cambridge University Press, New York.

[8] Frost, J. (2013). What Are the Effects of Multicollinearity and When Can I Ignore Them? Acedido a 5 de Dezembro de 2016 em: http://blog.minitab.com/blog/adventures-in-statistics/what-are-the-effects-of-multicollinearity-and-when-can-i-ignore-them

[9] [Moro et al., 2014] Moro, S., Cortez, P. e Rita, P. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014. Acedido a 22 de Dezembro de 2016 em: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

[10] Mathier, S. (2004-2016). Free Excel Training Course. Acedido várias vezes durante o preiodo de realização do projeto, em: http://www.excel-pratique.com/en/course.php.

[11] Pestana, D., Velosa, S (2010). Introdução à Probabilidade e à Estatística Volume I. 4ª Edição, Fundação Calouste Gulbenkian, Lisboa.

[12] Sen, A. e Srivastava, M. (1997). Regression Analysis: Theory, Methods, and Applications. 4ª edição, Springer Science & Business Media, New York.

[13] Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.

[14] Zaiontz, C. (2013). Significance Testing of the Logistic Regression Coefficients. Acedido a 15 de Dezembro de 2016 em: http://www.real-statistics.com/logistic-regression/significance-testing-logistic-regression-coefficients/

ANEXOS

9. COMUM

```
Sub btn_Caract_Amostrais_Click()
  F1_Caract_Amostrais.Show
End Sub

Sub btn_Comp_Variancia_Click()
  F2_Comp_Variancia.Show
End Sub

Sub btn_Infl_Variancias_Click()
  F3_Infl_Variancias.Show
End Sub

Sub btn_Reg_Lin_Click()
  F4_Regressao_Linear.Show
End Sub

Sub btn_Reg_Log_Click()
  F5_Regressao_Logistica.Show
End Sub
```

10. FORMULÁRIOS PARA INPUT

10.1. F1_CARACT_AMOSTRAIS

```
Private Sub btn_OK_CA_Click()
  If dados_CA.Value = "" Then
    dados_vazio_CA.Visible = True
    Exit Sub
  Else
    dados_vazio_CA.Visible = False
  End If

  Dim addr_oX As String, oX As Range

  addr_oX = dados_CA.Value
  Set oX = Range(addr_oX)

  Call Caract_Amostrais(oX)
  Unload Me
End Sub

Private Sub btn_cancel_CA_Click()
  Unload Me
End Sub
```

10.2. F2_COMP_VARIANCIA

```
Private Sub btn_OK_CV_Click()
  If dados_CV.Value = "" Then
    dados_vazio_CV.Visible = True
    Exit Sub
  Else
    dados_vazio_CV.Visible = False
  End If

  Dim addr_X As String, X As Range
  addr_X = dados_CV.Value
  Set X = Range(addr_X)
  Call Comp_Variancia(X)

  Unload Me
End Sub

Private Sub btn_cancel_CA_Click()
  Unload Me
End Sub
```

10.3. F3_INFL_VARIANCIAS

```
Private Sub btn_OK_IV_Click()
  If dados_IV.Value = "" Then
    dados_vazio_IV.Visible = True
    Exit Sub
  Else
    dados_vazio_IV.Visible = False
  End If

  If R2_IV.Value = "" Then
    R2_vazio_IV.Visible = True
    Entre_0_1_VIF.Visible = False
    Exit Sub
  ElseIf R2_IV.Value <= 0 Or R2_IV.Value >= 1 Then
```

```
    Entre_0_1_VIF.Visible = True
    R2_vazio_IV.Visible = False
    Exit Sub
  Else
    R2_vazio_IV.Visible = False
    Entre_0_1_VIF.Visible = False
  End If

  Dim addr_X As String, X As Range, R2 As Double
  addr_X = dados_IV.Value
  Set X = Range(addr_X)

  R2 = R2_IV.Value
  Call Infl_Variancia(X, R2)

  Unload Me
End Sub

Private Sub btn_cancel_IV_Click()
  Unload Me
End Sub
```

10.4. F4_REGRESSAO_LINEAR

```
Private Sub btn_OK_lin_Click()
  If Y_lin.Value = "" Then
    Y_vazio_lin.Visible = True
    Exit Sub
  Else
    Y_vazio_lin.Visible = False
  End If

  If X_lin.Value = "" Then
    X_vazio_lin.Visible = True
    Exit Sub
  Else
    X_vazio_lin.Visible = False
  End If

  If const1_lin.Value = False And const0_lin.Value = False
Then
    ct_vazio_lin.Visible = True
    Exit Sub
  Else
    ct_vazio_lin.Visible = False
  End If

  If alfa_lin.Value = "" Then
    alfa_vazio_lin.Visible = True
    Entre_0_1_in_lin.Visible = False
    Exit Sub
  ElseIf alfa_lin.Value <= 0 Or alfa_lin.Value >= 1 Then
    Entre_0_1_in_lin.Visible = True
    alfa_vazio_lin.Visible = False
    Exit Sub
  Else
    alfa_vazio_lin.Visible = False
    Entre_0_1_in_lin.Visible = False
  End If

  If Backward_lin.Value = False And Forward_lin.Value = False
And Stepwise_lin.Value = False Then
    met_vazio_lin.Visible = True
    Exit Sub
  Else
    met_vazio_lin.Visible = False
  End If

  If Stepwise_lin.Value = True Then
    If alfa_out_lin.Value = "" Then
      alfa_out_vazio_lin.Visible = True
      Exit Sub
    ElseIf alfa_out_lin.Value <= 0 Or alfa_out_lin.Value >= 1
Then
      Entre_0_1_out_lin.Visible = True
      Exit Sub
    Else
      alfa_out_vazio_lin.Visible = False
      Entre_0_1_out_lin.Visible = False
    End If
  End If

  Dim addr_oY As String, oY As Range, addr_oX As String, oX
As Range, ct As Integer, alfa As Double, alfa_out As Double

  addr_oY = Y_lin.Value
  Set oY = Range(addr_oY)
```

```
  addr_oX = X_lin.Value
  Set oX = Range(addr_oX)

  If const1_lin.Value Then
    ct = 1
  ElseIf const0_lin.Value Then
    ct = 0
  End If

  alfa = alfa_lin.Value
  If Backward_lin.Value = True Then
    Call Lin_Backward(oY, oX, ct, alfa)
  End If

  If Forward_lin.Value = True Then
    Call Lin_Forward(oY, oX, ct, alfa)
  End If

  If Stepwise_lin.Value = True Then
    alfa_out = alfa_out_lin.Value
    Call Lin_Stepwise(oY, oX, ct, alfa, alfa_out)
  End If

  Unload Me
End Sub

Private Sub btn_Cancel_lin_Click()
Unload Me
End Sub
```

10.5. F5_REGRESSAO_LOGISTICA

```
Private Sub btn_OK_log_Click()
  If Y_log.Value = "" Then
    Y_vazio_log.Visible = True
    Exit Sub
  Else
    Y_vazio_log.Visible = False
  End If

  If X_log.Value = "" Then
    X_vazio_log.Visible = True
    Exit Sub
  Else
    X_vazio_log.Visible = False
  End If

  If alfa_log.Value = "" Then
    alfa_vazio_log.Visible = True
    Entre_0_1_in_log.Visible = False
    Exit Sub
  ElseIf alfa_log.Value <= 0 Or alfa_log.Value >= 1 Then
    Entre_0_1_in_log.Visible = True
    alfa_vazio_log.Visible = False
    Exit Sub
  Else
    alfa_vazio_log.Visible = False
    Entre_0_1_in_log.Visible = False
  End If

 If pc_log.Value = "" Or pc_log.Value < 10 Then
    pc_vazio_log.Visible = True
    Exit Sub
  Else
    pc_vazio_log.Visible = False
  End If

  If Backward_log.Value = False And Forward_log.Value = False
And Stepwise_log.Value = False Then
    met_vazio_log.Visible = True
    Exit Sub
  Else
    met_vazio_log.Visible = False
  End If
    If Stepwise_log.Value = True Then
    If alfa_out_log.Value = "" Then
      alfa_out_vazio_log.Visible = True
      Exit Sub
    ElseIf alfa_out_log.Value <= 0 Or alfa_out_log.Value >= 1
Then
      Entre_0_1_out_log.Visible = True
      Exit Sub
    Else
      alfa_out_vazio_log.Visible = False
      Entre_0_1_out_log.Visible = False
    End If
  End If
```

```
  Dim addr_oY As String, oY As Range, addr_oX As String, oX
As Range, alfa As Double, alfa_out As Double, pc As Long
  addr_oY = Y_log.Value
  Set oY = Range(addr_oY)

  addr_oX = X_log.Value
  Set oX = Range(addr_oX)

  alfa = alfa_log.Value
  pc = pc_log.Value

  If Backward_log.Value = True Then
    Call Log_Backward(oY, oX, alfa, pc)
  End If

  If Forward_log.Value = True Then
    Call Log_Forward(oY, oX, alfa, pc)
  End If

  If Stepwise_log.Value = True Then
    alfa_out = alfa_out_log.Value
    Call Log_Stepwise(oY, oX, alfa, alfa_out, pc)
  End If
    Unload Me
End Sub

Private Sub btn_Cancel_log_Click()
  Unload Me
End Sub
```

11. PROGRAMAS

11.1. SUB_01_CARACT_AMOSTRAIS

```
    Sub Caract_Amostrais(amostra As Range)
  Folha_Dados = ActiveSheet.Name

  Sheets.Add after:=Sheets(Sheets.Count)
  Sheets(ActiveSheet.Name).Name = "Out_Caracteristicas"
  ActiveWindow.DisplayGridlines = False

  Worksheets(Folha_Dados).Activate
  Set output = Sheets("Out_Caracteristicas").Range("B2")

  Sheets("Out_Caracteristicas").Cells.HorizontalAlignment =
xlCenter
  Sheets("Out_Caracteristicas").Cells.VerticalAlignment =
xlCenter
  With Sheets("Out_Caracteristicas").Cells.Font
      .Name = "Times New Roman"
      .Size = 10
  End With
'-------------------------------------------------------------

  Dim cl As Long, rw As Long
  cl = amostra.Columns.Count
  rw = amostra.Rows.Count - 1

  ReDim h(1 To cl, 1 To 1)
  For Lin = 1 To cl
    h(Lin, 1) = amostra(1, Lin)
  Next Lin

  output.Cells(1, 1).Value = "Características Amostrais"
  Range(output.Cells(1, 1), output.Cells(1, 10)).Merge
  Range(output.Cells(1, 1), output.Cells(1, 10)).Font.Bold =
True
  Range(output.Cells(1, 1), output.Cells(1,
10)).Interior.Color = RGB(218, 238, 243)

  Range(output.Cells(4, 1), output.Cells(3 + cl, 1)) = h
  Range(output.Cells(4, 1), output.Cells(3 + cl,
1)).Font.Bold = True

  output.Cells(3, 2).Value = "Registos"
  output.Cells(3, 3).Value = "Observações"
  output.Cells(3, 4).Value = "% Missings"
  output.Cells(3, 5).Value = "Média"
  output.Cells(3, 6).Value = "Variância"
  output.Cells(3, 7).Value = "Desvio-Padrão"
  output.Cells(3, 8).Value = "Mediana"
  output.Cells(3, 9).Value = "Mínimo"
  output.Cells(3, 10).Value = "Máximo"
  output.Cells(3, 12).Value = "Seleção:"

  Range(output.Cells(3, 1), output.Cells(3, 12)).Font.Bold =
True
```

```
  For Lin = 1 To cl

    ReDim amostra_cl(1 To rw, 1 To 1)
    Dim R_amostra_cl As Range
    For lin1 = 1 To rw
      amostra_cl(lin1, 1) = amostra(lin1 + 1, Lin)
    Next lin1

    'Registos:
    output.Cells(Lin + 3, 2).Value =
WorksheetFunction.CountA(amostra_cl)
    output.Cells(Lin + 3, 2).NumberFormat = "#,##0"

    'Observacoes:
    output.Cells(Lin + 3, 3).Value =
WorksheetFunction.Count(amostra_cl)
    output.Cells(Lin + 3, 3).NumberFormat = "#,##0"

    '% Missings:
    output.Cells(Lin + 3, 4).Value =
(WorksheetFunction.CountA(amostra_cl) -
WorksheetFunction.Count(amostra_cl)) /
WorksheetFunction.CountA(amostra_cl)
    output.Cells(Lin + 3, 4).NumberFormat = "0.00%"

    'Media:
    output.Cells(Lin + 3, 5).Value =
WorksheetFunction.Average(amostra_cl)
    output.Cells(Lin + 3, 5).NumberFormat = "#,##0.00"

    'Variancia:
    output.Cells(Lin + 3, 6).Value =
WorksheetFunction.Var_S(amostra_cl)
    output.Cells(Lin + 3, 6).NumberFormat = "#,##0.00"

    'Desvio-Padrao:
    output.Cells(Lin + 3, 7).Value =
WorksheetFunction.StDev_S(amostra_cl)
    output.Cells(Lin + 3, 7).NumberFormat = "#,##0.00"

    'Mediana:
    output.Cells(Lin + 3, 8).Value =
WorksheetFunction.Median(amostra_cl)
    output.Cells(Lin + 3, 8).NumberFormat = "#,##0.00"

    'Minimo:
    output.Cells(Lin + 3, 9).Value =
WorksheetFunction.Min(amostra_cl)
    output.Cells(Lin + 3, 9).NumberFormat = "#,##0.00"

    'Maximo:
    output.Cells(Lin + 3, 10).Value =
WorksheetFunction.Max(amostra_cl)
    output.Cells(Lin + 3, 10).NumberFormat = "#,##0.00"

  Next Lin

  With Range(output.Cells(3, 1), output.Cells(Lin + 2,
10)).Borders
      .LineStyle = xlContinuous
      .Color = RGB(166, 166, 166)
  End With

  Range(output.Cells(3, 1), output.Cells(Lin + 2,
10)).BorderAround (xlDouble)
  output.Cells(3, 1).BorderAround (xlDouble)
  Range(output.Cells(4, 2), output.Cells(Lin + 2,
10)).BorderAround (xlDouble)

  output.Cells(4, 12).Value = amostra.Worksheet.Name
  output.Cells(5, 12).Value = amostra.Address(False, False)

  Sheets("Out_Caracteristicas").Range(output.Cells(4, 1),
output.Cells(3 + cl, 1)).HorizontalAlignment = xlLeft

  Sheets("Out_Caracteristicas").Cells.EntireColumn.AutoFit
  Sheets("Out_Caracteristicas").Cells.EntireRow.AutoFit

End Sub
```

11.2. SUB_02_COMP_VARIANCIA

```
Sub Comp_Variancia(X As Range)
  Folha_Dados = ActiveSheet.Name

  Sheets.Add after:=Sheets(Sheets.Count)
  Sheets(ActiveSheet.Name).Name = "Out_Comp_Variancia"
  ActiveWindow.DisplayGridlines = False

  Worksheets(Folha_Dados).Activate
  Set output = Sheets("Out_Comp_Variancia").Range("B2")

  With Sheets("Out_Comp_Variancia").Cells.Font
      .Name = "Times New Roman"
      .Size = 10
  End With

  Sheets("Out_Comp_Variancia").Cells.HorizontalAlignment =
xlCenter
  Sheets("Out_Comp_Variancia").Cells.VerticalAlignment =
xlCenter
'-------------------------------------------------------------
  Dim X_lin As Long, X_col As Long

  X_lin = X.Rows.Count - 1
  X_col = X.Columns.Count + 1

  output.Cells(1, 1).Value = "Componentes da Variância"
  Range(output.Cells(1, 1), output.Cells(1, X_col + 2)).Merge
  Range(output.Cells(1, 1), output.Cells(1, X_col +
2)).Font.Bold = True
  Range(output.Cells(1, 1), output.Cells(1, X_col +
2)).Interior.Color = RGB(218, 238, 243)

  ReDim X_const(1 To X_lin, 1 To X_col)
  For Lin = 1 To X_lin
      X_const(Lin, 1) = 1
    For col = 2 To X_col
      X_const(Lin, col) = X(Lin + 1, col - 1)
    Next col
  Next Lin

  ReDim X_quad(1 To X_lin, 1 To X_col)
  For Lin = 1 To X_lin
    For col = 1 To X_col
      X_quad(Lin, col) = WorksheetFunction.Power(X_const(Lin,
col), 2)
    Next col
  Next Lin

  ReDim X_quad_sum(1 To 1, 1 To X_col), X_quad_sum_raiz(1 To
1, 1 To X_col)
  For col = 1 To X_col
    For Lin = 1 To X_lin
      X_quad_sum(1, col) = X_quad_sum(1, col) + X_quad(Lin,
col)
    Next Lin
    X_quad_sum_raiz(1, col) = X_quad_sum(1, col) ^ (1 / 2)
  Next col

  ReDim X_aux(1 To X_lin, 1 To X_col)
  For Lin = 1 To X_lin
    For col = 1 To X_col
      X_aux(Lin, col) = X_const(Lin, col) /
X_quad_sum_raiz(1, col)
    Next col
  Next Lin

  ReDim inv_XTX(1 To X_col, 1 To X_col)
  inv_XTX =
WorksheetFunction.MMult(WorksheetFunction.Transpose(X_aux),
X_aux)


  ReDim A(1 To X_col, 1 To X_col)
  For Lin = 1 To X_col
    For col = 1 To X_col
      A(Lin, col) = inv_XTX(Lin, col)
    Next col
  Next Lin

  Dim nmaxi As Long, Delta As Single
  Delta = 1E-16
  nmaxi = 30

  'Calculo
  Pi = Application.Pi()
  ReDim U(1 To X_col, 1 To X_col)
```

```
  For Lin = 1 To X_col
    U(Lin, Lin) = 1
    For col = Lin + 1 To X_col
      U(Lin, col) = 0
      U(col, Lin) = 0
    Next col
  Next Lin

  ReDim r(1 To X_col, 1 To X_col), AuxT(1 To X_col, 1 To
X_col), AuxP(1 To X_col, 1 To X_col)
  iter = 0
  Do
    iter = iter + 1
    flag = 0
    For Lin = 1 To X_col - 1
      For col = Lin + 1 To X_col
        If Abs(A(Lin, col)) > Delta Then
          flag = 1

          If Abs(A(Lin, Lin) - A(col, col)) > Delta Then
            teta = 0.5 * Atn(2 * A(Lin, col) / (A(Lin, Lin) -
A(col, col)))
            If Abs(teta) > (Pi / 4) Then teta = teta - (Pi /
2) * Sgn(teta)
            Else
                If A(Lin, col) > 0 Then teta = Pi / 4 Else
teta = -Pi / 4
            End If

            For lin2 = 1 To X_col
                r(lin2, lin2) = 1
              For col2 = lin2 + 1 To X_col
                r(lin2, col2) = 0
                r(col2, lin2) = 0
              Next col2
            Next lin2

            r(Lin, Lin) = Cos(teta)
            r(col, col) = Cos(teta)
            r(Lin, col) = -Sin(teta)
            r(col, Lin) = Sin(teta)

            AuxT = WorksheetFunction.Transpose(r)
            AuxP = WorksheetFunction.MMult(AuxT, A)
            A = WorksheetFunction.MMult(AuxP, r)
            AuxP = WorksheetFunction.MMult(U, r)

            For lin2 = 1 To X_col
              For col2 = 1 To X_col
                U(lin2, col2) = AuxP(lin2, col2)
              Next col2
            Next lin2

          End If
        Next col
      Next Lin

    Loop Until flag = 0 Or iter > nmaxi

    If flag <> 1 Then
      'Construir matriz das componentes da variancia
      ReDim Somas(1 To X_col), Comps(1 To X_col, 1 To X_col)
      For Lin = 1 To X_col
        For col = 1 To X_col
          Comps(Lin, col) = WorksheetFunction.Power(U(Lin,
col), 2) / A(col, col)
        Next col

          Somas(Lin) = 0
        For col = 1 To X_col
          Somas(Lin) = Somas(Lin) + Comps(Lin, col)
        Next col
      Next Lin

      For Lin = 1 To X_col
        For col = 1 To X_col
          Comps(Lin, col) = Comps(Lin, col) / Somas(Lin)
        Next col
      Next Lin

      ReDim result_aux(1 To X_col, 1 To X_col + 1)
      For Lin = 1 To X_col
          result_aux(Lin, 1) = A(Lin, Lin)
        For col = 2 To X_col + 1
          result_aux(Lin, col) = Comps(col - 1, Lin)
        Next col
      Next Lin
```

```
  End If

  'Numeros Condicao
  ReDim v_pp(1 To X_col, 1 To 1)
  Dim max_aux As Double
  For Lin = 1 To X_col
    v_pp(Lin, 1) = result_aux(Lin, 1)
  Next Lin
  max_aux = Application.WorksheetFunction.Max(v_pp)

  ReDim result(1 To X_col, 1 To X_col + 2)
  For Lin = 1 To X_col
      result(Lin, 1) = result_aux(Lin, 1)
      result(Lin, 2) = WorksheetFunction.Power(max_aux /
result_aux(Lin, 1), 1 / 2)
    For col = 3 To X_col + 2
      result(Lin, col) = result_aux(Lin, col - 1)
    Next col
  Next Lin

  output.Cells(3, 1).Value = "Valores" & Chr(10) & "Próprios"
  Range(output.Cells(3, 1), output.Cells(4, 1)).Merge

  output.Cells(3, 2).Value = "Números" & Chr(10) & "Condição"
  Range(output.Cells(3, 2), output.Cells(4, 2)).Merge

  output.Cells(3, 3).Value = "Proporções da variância do
coeficiente de regressão de"
  Range(output.Cells(3, 3), output.Cells(3, X_col + 2)).Merge
  output.Cells(4, 3).Value = "   Const   "

  For col = 1 To X_col - 1
    output.Cells(4, 3 + col).Value = "   " & X(1, col) & "
"
  Next col

  Range(output.Cells(3, 1), output.Cells(4, X_col +
2)).Font.Bold = True

  Range(output.Cells(5, 1), output.Cells(X_col + 4, X_col +
2)) = result
  Range(output.Cells(5, 1), output.Cells(X_col + 4, X_col +
2)).NumberFormat = "#,##0.00"

  With Range(output.Cells(5, 1), output.Cells(X_col + 4,
X_col + 2)).Borders
      .LineStyle = xlContinuous
      .Color = RGB(166, 166, 166)
  End With

  With Range(output.Cells(4, 1), output.Cells(4, X_col +
2)).Borders(xlInsideVertical)
      .LineStyle = xlContinuous
      .Color = RGB(166, 166, 166)
  End With
  Range(output.Cells(3, 1), output.Cells(X_col + 4, X_col +
2)).BorderAround (xlDouble)
  Range(output.Cells(3, 1), output.Cells(4, X_col +
2)).BorderAround (xlDouble)
  Range(output.Cells(3, 2), output.Cells(4, 2)).BorderAround
(xlDouble)
  Range(output.Cells(4, 2), output.Cells(X_col + 4,
2)).BorderAround (xlDouble)

  output.Cells(X_col + 6, 1).Value = "Seleção:"
  output.Cells(X_col + 7, 1).Value = X.Worksheet.Name
  output.Cells(X_col + 8, 1).Value = X.Address(False, False)
  output.Cells(X_col + 6, 1).Font.Bold = True

  Sheets("Out_Comp_Variancia").Cells.EntireColumn.AutoFit
  Sheets("Out_Comp_Variancia").Cells.EntireRow.AutoFit

End Sub
```

11.3. SUB_03_INFL_VARIANCIA

```
Sub Infl_Variancia(oX As Range, R2_min As Double)
  Folha_Dados = ActiveSheet.Name

  Sheets.Add after:=Sheets(Sheets.Count)
    Sheets(ActiveSheet.Name).Name = "Out_Infl_Variancia"
  ActiveWindow.DisplayGridlines = False

  Worksheets(Folha_Dados).Activate
  Set output = Sheets("Out_Infl_Variancia").Range("B2")

  With Sheets("Out_Infl_Variancia").Cells.Font
```

```
      .Name = "Times New Roman"
      .Size = 10
  End With

  Sheets("Out_Infl_Variancia").Cells.HorizontalAlignment =
xlCenter
  Sheets("Out_Infl_Variancia").Cells.VerticalAlignment =
xlCenter
'-------------------------------------------------------------

  Dim tol As Double
  tol = 1 / (1 - R2_min)

  Dim oX_col As Long, oX_lin As Long, X_col As Long
  oX_col = oX.Columns.Count
  oX_lin = oX.Rows.Count - 1

  ReDim h(1 To 1, 1 To oX_col)

  For col = 1 To oX_col
    h(1, col) = oX(1, col)
  Next col

  X_col = oX_col
  k = 1

  ReDim X(1 To oX_lin, 1 To oX_col) As Variant
  For col = 1 To oX_col
    For Lin = 1 To oX_lin
      X(Lin, col) = oX(Lin + 1, col)
    Next Lin
  Next col

  Call correlacoes(X, MCorrelacao_inv, max_diagonal,
ordem_max)
  If max_diagonal > tol Then

    Do Until max_diagonal <= tol Or k = oX_col + 2

    Call remove_col(X, ordem_max, X, col_X)
    Call remove_col(h, ordem_max, h, col_H)
    Call correlacoes(X, MCorrelacao_inv, max_diagonal,
ordem_max)
    k = k + 1
    Loop

  Else
  End If

  X_col = UBound(X, 2)
  output.Cells(1, 1) = "Fatores de Inflação das Variâncias"
  Range(output.Cells(1, 1), output.Cells(1, X_col + 1)).Merge
  Range(output.Cells(1, 1), output.Cells(1, X_col +
1)).Font.Bold = True
  Range(output.Cells(1, 1), output.Cells(1, X_col +
1)).Interior.Color = RGB(218, 238, 243)

  output.Cells(3, 1) = "Tolerância R2:"
  output.Cells(3, 1).Characters(Start:=13,
Length:=1).Font.Superscript = True
  output.Cells(3, 1).HorizontalAlignment = xlRight

  output.Cells(3, 3) = R2_min
  output.Cells(3, 3).NumberFormat = "#,##0.000"

  output.Cells(4, 1) = "Tolerância VIF:"
  output.Cells(4, 3) = tol
  output.Cells(4, 3).NumberFormat = "#,##0.000"
  output.Cells(4, 1).HorizontalAlignment = xlRight

  Range(output.Cells(3, 1), output.Cells(3, 2)).Merge
  Range(output.Cells(3, 1), output.Cells(3, 2)).Font.Bold =
True

  Range(output.Cells(4, 1), output.Cells(4, 2)).Merge
  Range(output.Cells(4, 1), output.Cells(4, 2)).Font.Bold =
True

  Range(output.Cells(6, 2), output.Cells(6, X_col + 1)) = h
  Range(output.Cells(6, 2), output.Cells(6, X_col +
1)).Font.Bold = True

  Range(output.Cells(7, 1), output.Cells(6 + X_col, 1)) =
WorksheetFunction.Transpose(h)
  Range(output.Cells(7, 1), output.Cells(7 + X_col + 1,
1)).Font.Bold = True
```

```
  output.Cells(6 + X_col + 2, 1).Value = "Seleção:"
  output.Cells(6 + X_col + 3, 1).Value = oX.Worksheet.Name
  output.Cells(6 + X_col + 4, 1).Value = oX.Address(False,
False)
  output.Cells(6 + X_col + 2, 1).Font.Bold = True

  Range(output.Cells(7, 2), output.Cells(X_col + 6, X_col +
1)) = MCorrelacao_inv
  Range(output.Cells(7, 2), output.Cells(X_col + 6, X_col +
1)).NumberFormat = "#,##0.000"

  With Range(output.Cells(6, 1), output.Cells(6 + X_col,
X_col + 1)).Borders
      .LineStyle = xlContinuous
      .Color = RGB(166, 166, 166)
  End With
  Range(output.Cells(6, 1), output.Cells(6 + X_col, X_col +
1)).BorderAround (xlDouble)
  output.Cells(6, 1).BorderAround (xlDouble)
  Range(output.Cells(7, 2), output.Cells(6 + X_col, X_col +
1)).BorderAround (xlDouble)

  For Lin = 2 To X_col + 1
    output.Cells(5 + Lin, Lin).Interior.Color = RGB(218, 238,
243)
  Next Lin

  Sheets("Out_Infl_Variancia").Cells.EntireColumn.AutoFit
  Sheets("Out_Infl_Variancia").Cells.EntireRow.AutoFit


End Sub
```

11.4. SUB_04_LIN_BACKWARDS

```
Sub Lin_Backward(oY As Range, oX As Range, ct As Integer,
alfa As Double)
  Folha_Dados = ActiveSheet.Name

  Sheets.Add after:=Sheets(Sheets.Count)
  Sheets(ActiveSheet.Name).Name = "Out_Backward_Linear"
  ActiveWindow.DisplayGridlines = False

  Worksheets(Folha_Dados).Activate

  Set output = Sheets("Out_Backward_Linear").Range("B4")

  With Sheets("Out_Backward_Linear").Cells.Font
      .Name = "Times New Roman"
      .Size = 10
  End With

  Sheets("Out_Backward_Linear").Cells.HorizontalAlignment =
xlCenter
  Sheets("Out_Backward_Linear").Cells.VerticalAlignment =
xlCenter
'-------------------------------------------------------------


  Dim oX_col As Long, oX_lin As Long
  oX_col = oX.Columns.Count
  oX_lin = oX.Rows.Count - 1

  output.Cells(-1, 1).Value = "Regressão Linear: Método de
Seleção Regressiva (Backward Selection)"
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Merge
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Font.Bold = True
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Interior.Color = RGB(218, 238, 243)

  'Para o layout do quadro de p-values a escrever no excel
  'oH: Header do quadro
  'H_X: Header com nomes das variaveis a usar no index-match
  'Q_pv: vetor de p-values
  ReDim oH(1 To 1, 1 To oX_col + 1), H_X(1 To 1, 1 To oX_col
+ 1), Q_pv(1 To 1, 1 To oX_col + 2)

  For col = 1 To oX_col
    oH(1, col) = oX(1, col)
    H_X(1, col) = oX(1, col)
  Next col
    oH(1, oX_col + 1) = "const"
    H_X(1, oX_col + 1) = "const"

  Range(output.Cells(1, 2), output.Cells(1, oX_col + 2)) = oH
  output.Cells(1, oX_col + 4) = "R2"
  output.Cells(1, oX_col + 4).Characters(Start:=2,
Length:=1).Font.Superscript = True
```

```
  Range(output.Cells(1, 2), output.Cells(1, oX_col +
4)).Font.Bold = True

  'Parametros para a primeira iteracao:
  'consideramos a matriz de planeamento que o utilizador
seleccionou
  aux_max = 0
  max_P_V = 999

  ReDim X(1 To oX_lin, 1 To oX_col) As Variant
  For col = 1 To oX_col
    For Lin = 1 To oX_lin
      X(Lin, col) = oX(Lin + 1, col)
    Next Lin
  Next col

  ReDim y(1 To oX_lin, 1 To 1) As Variant
  For Lin = 1 To oX_lin
    y(Lin, 1) = oY(Lin + 1, 1)
  Next Lin

  k = 1 'para iniciar as iteracoes

  Do Until max_P_V < alfa Or k = oX_col + 1 'ate todas as
variaveis serem significativas ou apos ter-se retirado todas
as variaveis

    'Para criar a nova matriz de planeamento
    Call remove_col(X, aux_max, X, col_X)
    X_col = UBound(X, 2)

    'construir o header correspondente das variaveis a
mostrar no quadro
    Call remove_col(H_X, aux_max, H_X, col_H)

    'Funcao Linest
    linest_output = Application.WorksheetFunction.LinEst(y,
X, ct, 1)

    'Para pegar no output do linest e calcular os P-Values
    Call P_Value_T(X_col, linest_output, ct, QF, P_V,
max_P_V, aux_max, PV_ult)

    'juntar os p-values no quadro a escrver na folha
    ReDim Q_pv(1 To 1, 1 To oX_col + 1)
    For col = 1 To oX_col + 1
      On Error Resume Next 'para ficar em branco quando o
index-match nao encontrar correspondencia
      Q_pv(1, col) = WorksheetFunction.Index(P_V, 1,
WorksheetFunction.Match(oH(1, col), H_X, 0))
    Next col

    'P-Values
    Range(output.Cells(k + 1, 2), output.Cells(k + 1, oX_col
+ 2)) = Q_pv

    'R2
    output.Cells(k + 1, oX_col + 4) =
Application.WorksheetFunction.Index(linest_output, 3, 1)

    'Formatacao do Quadro dos P-Values
    For col = 1 To oX_col + 2
      If output.Cells(k + 1, col).Value = max_P_V And max_P_V
>= alfa Then
        output.Cells(k + 1, col).Interior.Color = RGB(217,
217, 217)
      End If
    Next col

    k = k + 1
  Loop

  Range(output.Cells(2, 2), output.Cells(k, oX_col +
4)).NumberFormat = "#,##0.000"

  With Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).Borders
      .LineStyle = xlContinuous
      .Color = RGB(166, 166, 166)
  End With
  Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).BorderAround (xlDouble)
  Range(output.Cells(1, 2), output.Cells(1, oX_col +
2)).BorderAround (xlDouble)

  Range(output.Cells(1, oX_col + 4), output.Cells(k, oX_col +
4)).BorderAround (xlDouble)
  output.Cells(1, oX_col + 4).BorderAround (xlDouble)
```

```
  With Range(output.Cells(2, oX_col + 4), output.Cells(k,
oX_col + 4)).Borders(xlInsideHorizontal)
       .LineStyle = xlContinuous
       .Color = RGB(166, 166, 166)
  End With

  j = k

  If max_P_V >= alfa Then 'Nao temos modelo

    output.Cells(k + 3, 2) = "a"
    output.Cells(k + 3, 2).Cells.Font.Name = "Symbol"
    output.Cells(k + 3, 2).Font.Bold = True
    output.Cells(k + 3, 3) = alfa
    output.Cells(k + 3, 3).NumberFormat = "0.000"

    With Range(output.Cells(k + 3, 2), output.Cells(k + 3,
3)).Borders
         .LineStyle = xlContinuous
         .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(k + 3, 2), output.Cells(k + 3,
3)).BorderAround (xlDouble)

  Else

  'Quadros da ultima iteracao: informacao do modelo
seleccionado pelo algoritmo
    ReDim Stats1(1 To 3, 1 To 1)
    For Lin = 1 To 3
      Stats1(Lin, 1) =
Application.WorksheetFunction.Index(linest_output, Lin + 2,
1)
    Next Lin

    ReDim Stats2(1 To 3, 1 To 1)
    For Lin = 1 To 3
      Stats2(Lin, 1) =
Application.WorksheetFunction.Index(linest_output, Lin + 2,
2)
    Next Lin

    j = k + 3
    output.Cells(j + 0, 1) = "Coeficientes"
    output.Cells(j + 1, 1) = "EMQ"
    output.Cells(j + 2, 1) = "s(b)"
    output.Cells(j + 2, 1).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(j + 3, 1) = "Estat. Teste"
    output.Cells(j + 4, 1) = "P-Value"
    Range(output.Cells(j + 0, 2), output.Cells(j + 0, X_col +
2)) = H_X
    Range(output.Cells(j + 0, 1), output.Cells(j + 0, X_col +
2)).Font.Bold = True
    Range(output.Cells(j + 1, 1), output.Cells(j + 4,
1)).Font.Bold = True
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
2)) = QF
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
2)).NumberFormat = "#,##0.000"

    With Range(output.Cells(j + 0, 1), output.Cells(j + 4,
X_col + 2)).Borders
         .LineStyle = xlContinuous
         .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, 1), output.Cells(j + 4, X_col +
2)).BorderAround (xlDouble)
    output.Cells(j + 0, 1).BorderAround (xlDouble)
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
2)).BorderAround (xlDouble)

    output.Cells(j + 0, X_col + 4) = "a"
    output.Cells(j + 0, X_col + 4).Cells.Font.Name = "Symbol"
    output.Cells(j + 0, X_col + 4).Font.Bold = True
    output.Cells(j + 0, X_col + 5) = alfa
    output.Cells(j + 0, X_col + 5).NumberFormat = "0.000"

    With Range(output.Cells(j + 0, X_col + 4), output.Cells(j
+ 0, X_col + 5)).Borders
         .LineStyle = xlContinuous
         .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, X_col + 4), output.Cells(j + 0,
X_col + 5)).BorderAround (xlDouble)

    output.Cells(j + 2, X_col + 4) = "R2"
```

```
    output.Cells(j + 2, X_col + 4).Characters(Start:=2,
Length:=1).Font.Superscript = True

    output.Cells(j + 3, X_col + 4) = "Teste-F"
    output.Cells(j + 4, X_col + 4) = "SQReg"
    output.Cells(j + 4, X_col + 4).Characters(Start:=3,
Length:=3).Font.Subscript = True
    Range(output.Cells(j + 2, X_col + 5), output.Cells(j + 4,
X_col + 5)) = Stats1
    Range(output.Cells(j + 2, X_col + 5), output.Cells(j + 4,
X_col + 5)).NumberFormat = "#,##0.000"
    Range(output.Cells(j + 2, X_col + 4), output.Cells(j + 4,
X_col + 4)).Font.Bold = True

    With Range(output.Cells(j + 2, X_col + 4), output.Cells(j
+ 4, X_col + 5)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 2, X_col + 4), output.Cells(j + 4,
X_col + 5)).BorderAround (xlDouble)

    output.Cells(j + 2, X_col + 7) = "S"
    output.Cells(j + 3, X_col + 7) = "G.L."
    output.Cells(j + 4, X_col + 7) = "SQe"
    output.Cells(j + 4, X_col + 7).Characters(Start:=3,
Length:=1).Font.Subscript = True
    Range(output.Cells(j + 2, X_col + 8), output.Cells(j + 4,
X_col + 8)) = Stats2
    Range(output.Cells(j + 2, X_col + 8), output.Cells(j + 4,
X_col + 8)).NumberFormat = "#,##0.000"
    output.Cells(j + 3, X_col + 8).NumberFormat = "#,##0"
    Range(output.Cells(j + 2, X_col + 7), output.Cells(j + 4,
X_col + 7)).Font.Bold = True

    With Range(output.Cells(j + 2, X_col + 7), output.Cells(j
+ 4, X_col + 8)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 2, X_col + 7), output.Cells(j + 4,
X_col + 8)).BorderAround (xlDouble)

  End If

  output.Cells(j + 6, 1).Value = "Seleção Y:"
  output.Cells(j + 6, 1).Font.Bold = True
  output.Cells(j + 7, 1).Value = oY.Worksheet.Name
  output.Cells(j + 8, 1).Value = oY.Address(False, False)

  output.Cells(j + 10, 1).Value = "Seleção X:"
  output.Cells(j + 10, 1).Font.Bold = True
  output.Cells(j + 11, 1).Value = oX.Worksheet.Name
  output.Cells(j + 12, 1).Value = oX.Address(False, False)

  Sheets("Out_Backward_Linear").Cells.EntireColumn.AutoFit
  Sheets("Out_Backward_Linear").Cells.EntireRow.AutoFit

End Sub
```

11.5. SUB_05_LIN_FORWARD

```
Sub Lin_Forward(oY As Range, oX As Range, ct As Integer, alfa
As Double)
  Folha_Dados = ActiveSheet.Name

  Sheets.Add after:=Sheets(Sheets.Count)
  Sheets(ActiveSheet.Name).Name = "Out_Forward_Linear"
  ActiveWindow.DisplayGridlines = False

  Worksheets(Folha_Dados).Activate
  Set output = Sheets("Out_Forward_Linear").Range("B4")

  With Sheets("Out_Forward_Linear").Cells.Font
      .Name = "Times New Roman"
      .Size = 10
  End With

  Sheets("Out_Forward_Linear").Cells.HorizontalAlignment =
xlCenter
  Sheets("Out_Forward_Linear").Cells.VerticalAlignment =
xlCenter
'-------------------------------------------------------------

  Dim oX_col As Long, oX_lin As Long
  oX_col = oX.Columns.Count
  oX_lin = oX.Rows.Count - 1
```

```
  output.Cells(-1, 1).Value = "Regressão Linear: Método de
Seleção Progressiva (Forward Selection)"
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Merge
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Font.Bold = True
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Interior.Color = RGB(218, 238, 243)

  'Para o layout do quadro de p-values a escrever no excel
  'oH: Header do quadro
  'H_X: Header com nomes das variaveis a usar no index-match
  'Q_pv: vetor de p-values organizados
  ReDim oH(1 To 1, 1 To oX_col + 1), H_X(1 To 1, 1 To oX_col
+ 1)
  ReDim QF1(1 To oX_col + 2, 1 To oX_col + 1), QF2(1 To
oX_col + 2, 1 To oX_col + 1), QF3(1 To oX_col + 2, 1 To
oX_col + 1), HF(1 To oX_col + 2, 1 To oX_col + 1)
  ReDim QF4(1 To oX_col + 2, 1 To oX_col + 1), Stats1(1 To
oX_col + 2, 1 To 2), Stats2(1 To oX_col + 2, 1 To 2),
Stats3(1 To oX_col + 2, 1 To 2), pv_aux_final(1 To oX_col +
2, 1 To 1)

  For col = 1 To oX_col
    oH(1, col) = oX(1, col)
    H_X(1, col) = oX(1, col)
  Next col
    oH(1, oX_col + 1) = "const"
    H_X(1, oX_col + 1) = "const"

  Range(output.Cells(1, 2), output.Cells(1, oX_col + 2)) = oH
  output.Cells(1, oX_col + 4) = "R2"
  output.Cells(1, oX_col + 4).Characters(Start:=2,
Length:=1).Font.Superscript = True
  Range(output.Cells(1, 2), output.Cells(1, oX_col +
4)).Font.Bold = True

  'Parametros para a primeira iteracao:
  'consideramos a matriz de planeamento que o utilizador
seleccionou
  aux_max = 0
  PV_in = 999
  X_col = oX_col
```

```
  ReDim X(1 To oX_lin, 1 To oX_col) As Variant
  For col = 1 To oX_col
    For Lin = 1 To oX_lin
      X(Lin, col) = oX(Lin + 1, col)
    Next Lin
  Next col

  ReDim y(1 To oX_lin, 1 To 1) As Variant
  For Lin = 1 To oX_lin
    y(Lin, 1) = oY(Lin + 1, 1)
  Next Lin

  k = 1 'primeira iteracao
  ReDim R2_output(1 To 1, 1 To X_col)
  For col = 1 To X_col
    ReDim X_teste(1 To oX_lin, 1 To 1)
    For Lin = 1 To oX_lin
      X_teste(Lin, 1) = oX(Lin + 1, col)
    Next Lin

    linest_output = Application.WorksheetFunction.LinEst(y,
X_teste, ct, 1)

    R2 = Application.WorksheetFunction.Index(linest_output,
3, 1)
    R2_output(1, col) = R2
  Next col

  max_R2 = Application.WorksheetFunction.Max(R2_output)
  aux_max_r2 = Application.WorksheetFunction.Match(max_R2,
R2_output, 0)

  'Para criar a nova matriz de planeamento base, com menos
uma coluna
  'Matriz de Planeamento do modelo com a 1a variavel
seleccionada
  Call remove_col(X, aux_max_r2, X, col_X)
  Call remove_col(H_X, aux_max_r2, H_X, col_H)

  ReDim X_modelo(1 To oX_lin, 1 To 1)
  For Lin = 1 To oX_lin
    X_modelo(Lin, 1) = col_X(Lin, 1)
  Next Lin
```

```
  ReDim H_modelo(1 To 1, 1 To 2)
  H_modelo(1, 1) = col_H(1, 1)
  H_modelo(1, 2) = "const"

  For col = 1 To 2
    HF(k, col) = H_modelo(1, col)
  Next col

  'Funcao Linest
  linest_output = Application.WorksheetFunction.LinEst(y,
X_modelo, ct, 1)

  'Para pegar no output do linest e calcular os P-Values
  Call P_Value_T(1, linest_output, ct, QF, P_V, max_P_V,
aux_max, PV_in)

  'juntar os p-values no quadro a escrver na folha
  ReDim Q_pv(1 To 1, 1 To oX_col + 1)
  For col = 1 To oX_col + 1
    On Error Resume Next 'para ficar em branco quando o
index-match nao encontrar correspondencia
    Q_pv(1, col) = WorksheetFunction.Index(P_V, 1,
WorksheetFunction.Match(oH(1, col), H_modelo, 0))
  Next col

  'P-Values
  Range(output.Cells(k + 1, 2), output.Cells(k + 1, oX_col +
2)) = Q_pv
  'R2
  output.Cells(k + 1, oX_col + 4) =
Application.WorksheetFunction.Index(linest_output, 3, 1)
  Range(output.Cells(k + 1, 2), output.Cells(k + 1, oX_col +
4)).NumberFormat = "#,##0.000"

  'O modelo final sera o obtido na penultima iteracao
  'informacao do modelo seleccionado a cada iteracao
   For col = 1 To 2
    QF1(k, col) = QF(1, col)
    QF2(k, col) = QF(2, col)
    QF3(k, col) = QF(3, col)
    QF4(k, col) = QF(4, col)
    Stats1(k, col) =
Application.WorksheetFunction.Index(linest_output, 3, col)
    Stats2(k, col) =
Application.WorksheetFunction.Index(linest_output, 4, col)
    Stats3(k, col) =
Application.WorksheetFunction.Index(linest_output, 5, col)
  Next col

  'Auxiliar para apurar a iteracao com o modelo seleccionado:
sera o ultimo em que todas as variaveis sao significativas
  If PV_in < alfa Then
    pv_aux_final(k, 1) = 1
  Else
    pv_aux_final(k, 1) = 0
  End If

  k = k + 1

  If PV_in >= alfa Then 'O algoritmo para e nao temos modelo

    Range(output.Cells(k, 2), output.Cells(k, oX_col +
2)).Interior.Color = RGB(217, 217, 217)
    output.Cells(k, oX_col + 4).Interior.Color = RGB(217,
217, 217)
    For col = 1 To oX_col + 2
      If output.Cells(k, col).Value = PV_in And PV_in >= alfa
Then
        output.Cells(k, col).Interior.Color = RGB(166, 166,
166)
      End If
    Next col

    With Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).BorderAround (xlDouble)
    Range(output.Cells(1, 2), output.Cells(1, oX_col +
2)).BorderAround (xlDouble)

    Range(output.Cells(1, oX_col + 4), output.Cells(k, oX_col
+ 4)).BorderAround (xlDouble)
    output.Cells(1, oX_col + 4).BorderAround (xlDouble)
```

```
    With Range(output.Cells(2, oX_col + 4), output.Cells(k,
oX_col + 4)).Borders(xlInsideHorizontal)
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With

    output.Cells(4, 2) = "a"
    output.Cells(4, 2).Cells.Font.Name = "Symbol"
    output.Cells(4, 2).Font.Bold = True
    output.Cells(4, 3) = alfa
    output.Cells(4, 3).NumberFormat = "0.000"

    With Range(output.Cells(4, 2), output.Cells(4,
3)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(4, 2), output.Cells(4,
3)).BorderAround (xlDouble)

  Else 'continuamos o algoritmo normalmente
    Do Until PV_in >= alfa Or k = oX_col + 1

      For col = 1 To X_col
        Call remove_col(X, col, X_temp, col_X)
        Call add_col(X_modelo, col_X, X_teste)

        linest_output =
Application.WorksheetFunction.LinEst(y, X_teste, ct, 1)

        R2 =
Application.WorksheetFunction.Index(linest_output, 3, 1)
        R2_output(1, col) = R2
      Next col

      max_R2 = Application.WorksheetFunction.Max(R2_output)
      aux_max_r2 =
Application.WorksheetFunction.Match(max_R2, R2_output, 0)

      'Matriz de planeamento base, com menos uma coluna
      'Matriz de Planeamento com a variavel seleccionada
      If UBound(X, 2) = 1 Then
        For Lin = 1 To oX_lin
          col_X(Lin, 1) = X(Lin, 1)
        Next Lin
        X_col = 0
      Else
        Call remove_col(X, aux_max_r2, X, col_X)
        X_col = UBound(X, 2)
      End If

      Call add_col(X_modelo, col_X, X_modelo)
      X_modelo_col = UBound(X_modelo, 2)

      Call remove_col(H_X, aux_max_r2, H_X, col_H)

      'Adicionar a coluna na penultima posicao
      ReDim H_modelo1(1 To 1, 1 To k + 1)
      For col = 1 To k - 1
        H_modelo1(1, col) = H_modelo(1, col)
      Next col
        H_modelo1(1, k) = col_H(1, 1)
        H_modelo1(1, k + 1) = "const"

      ReDim H_modelo(1 To 1, 1 To k + 1)
      For col = 1 To k + 1
        H_modelo(1, col) = H_modelo1(1, col)
      Next col

      'Funcao Linest
      linest_output = Application.WorksheetFunction.LinEst(y,
X_modelo, ct, 1)

      'Para pegar no output do linest e calcular os P-Values
      Call P_Value_T(X_modelo_col, linest_output, ct, QF,
P_V, max_P_V, aux_max, PV_in)

      'juntar os p-values no quadro a escrver na folha
      ReDim Q_pv(1 To 1, 1 To oX_col + 1)
      For col = 1 To oX_col + 1
        On Error Resume Next 'para ficar em branco quando o
index-match nao encontrar correspondencia
        Q_pv(1, col) = WorksheetFunction.Index(P_V, 1,
WorksheetFunction.Match(oH(1, col), H_modelo, 0))
      Next col

      'P-Values
```

```
      Range(output.Cells(k + 1, 2), output.Cells(k + 1,
oX_col + 2)) = Q_pv
      'R2
      output.Cells(k + 1, oX_col + 4) =
Application.WorksheetFunction.Index(linest_output, 3, 1)
      Range(output.Cells(k + 1, 2), output.Cells(k + 1,
oX_col + 4)).NumberFormat = "#,##0.000"


      'O modelo final sera o obtido na penultima iteracao
      'informacao do modelo seleccionado a cada iteracao
      For col = 1 To X_modelo_col + 1
        HF(k, col) = H_modelo(1, col)
      Next col

      For col = 1 To X_modelo_col + 1
        QF1(k, col) = QF(1, col)
        QF2(k, col) = QF(2, col)
        QF3(k, col) = QF(3, col)
        QF4(k, col) = QF(4, col)
      Next col

      For col = 1 To 2
        Stats1(k, col) =
Application.WorksheetFunction.Index(linest_output, 3, col)
        Stats2(k, col) =
Application.WorksheetFunction.Index(linest_output, 4, col)
        Stats3(k, col) =
Application.WorksheetFunction.Index(linest_output, 5, col)
      Next col

      'Auxiliar para apurar a iteracao com o modelo
seleccionado: sera o ultimo em que todas as variaveis sao
significativas
      If PV_in < alfa Then
        pv_aux_final(k, 1) = 1
      Else
        pv_aux_final(k, 1) = 0
      End If

      k = k + 1
    Loop
```

```
    Range(output.Cells(k, 2), output.Cells(k, oX_col +
2)).Interior.Color = RGB(217, 217, 217)
    output.Cells(k, oX_col + 4).Interior.Color = RGB(217,
217, 217)
    For col = 1 To oX_col + 2
      If output.Cells(k, col).Value = PV_in And PV_in >= alfa
Then
        output.Cells(k, col).Interior.Color = RGB(166, 166,
166)
      End If
    Next col

    With Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).BorderAround (xlDouble)
    Range(output.Cells(1, 2), output.Cells(1, oX_col +
2)).BorderAround (xlDouble)

    Range(output.Cells(1, oX_col + 4), output.Cells(k, oX_col
+ 4)).BorderAround (xlDouble)
    output.Cells(1, oX_col + 4).BorderAround (xlDouble)
    With Range(output.Cells(2, oX_col + 4), output.Cells(k,
oX_col + 4)).Borders(xlInsideHorizontal)
    .LineStyle = xlContinuous
    .Color = RGB(166, 166, 166)
    End With

    'Temos que determinar em que iteracao foi testado o
modelo seleccionado
    iter = 0
    Lin = UBound(pv_aux_final, 1)
    Do Until pv_aux_final(Lin, 1) = 1
      Lin = Lin - 1
      iter = Lin
    Loop

    ReDim H_X(1 To 1, 1 To k + 1)
    For col = 1 To k + 1
      H_X(1, col) = HF(iter, col)
    Next col
```

```
    X_modelo_col =
Application.WorksheetFunction.Match("const", H_X, 0)

    ReDim QF_final(1 To 4, 1 To X_modelo_col + 2)
    For col = 1 To X_modelo_col + 2
      QF_final(1, col) = QF1(iter, col)
      QF_final(2, col) = QF2(iter, col)
      QF_final(3, col) = QF3(iter, col)
      QF_final(4, col) = QF4(iter, col)
    Next col

    ReDim Stats_final1(1 To 3, 1 To 1)
    Stats_final1(1, 1) = Stats1(iter, 1)
    Stats_final1(2, 1) = Stats2(iter, 1)
    Stats_final1(3, 1) = Stats3(iter, 1)

    ReDim Stats_final2(1 To 3, 1 To 1)
    Stats_final2(1, 1) = Stats1(iter, 2)
    Stats_final2(2, 1) = Stats2(iter, 2)
    Stats_final2(3, 1) = Stats3(iter, 2)

    j = k + 3
    X_col = X_modelo_col

    output.Cells(j + 0, 1) = "Coeficientes"
    output.Cells(j + 1, 1) = "EMQ"
    output.Cells(j + 2, 1) = "s(b)"
    output.Cells(j + 2, 1).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(j + 3, 1) = "Estat. Teste"
    output.Cells(j + 4, 1) = "P-Value"
    Range(output.Cells(j + 0, 2), output.Cells(j + 0, X_col +
1)) = H_X
    Range(output.Cells(j + 0, 1), output.Cells(j + 0, X_col +
3)).Font.Bold = True
    Range(output.Cells(j + 1, 1), output.Cells(j + 4,
1)).Font.Bold = True
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
3)) = QF_final
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
3)).NumberFormat = "#,##0.000"
    With Range(output.Cells(j + 0, 1), output.Cells(j + 4,
X_col + 1)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, 1), output.Cells(j + 4, X_col +
1)).BorderAround (xlDouble)
    output.Cells(j + 0, 1).BorderAround (xlDouble)
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
1)).BorderAround (xlDouble)

    output.Cells(j + 0, X_col + 3) = "a"
    output.Cells(j + 0, X_col + 3).Cells.Font.Name = "Symbol"
    output.Cells(j + 0, X_col + 3).Font.Bold = True
    output.Cells(j + 0, X_col + 4) = alfa
    output.Cells(j + 0, X_col + 4).NumberFormat = "0.000"

    With Range(output.Cells(j + 0, X_col + 3), output.Cells(j
+ 0, X_col + 4)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, X_col + 3), output.Cells(j + 0,
X_col + 4)).BorderAround (xlDouble)

    output.Cells(j + 2, X_col + 3) = "R2"
    output.Cells(j + 2, X_col + 3).Characters(Start:=2,
Length:=1).Font.Superscript = True

    output.Cells(j + 3, X_col + 3) = "Teste-F"
    output.Cells(j + 4, X_col + 3) = "SQReg"
    output.Cells(j + 4, X_col + 3).Characters(Start:=3,
Length:=3).Font.Subscript = True
    Range(output.Cells(j + 2, X_col + 4), output.Cells(j + 4,
X_col + 4)) = Stats_final1
    Range(output.Cells(j + 2, X_col + 4), output.Cells(j + 4,
X_col + 4)).NumberFormat = "#,##0.000"
    Range(output.Cells(j + 2, X_col + 3), output.Cells(j + 4,
X_col + 3)).Font.Bold = True

    With Range(output.Cells(j + 2, X_col + 3), output.Cells(j
+ 4, X_col + 4)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
```

```
    End With
    Range(output.Cells(j + 2, X_col + 3), output.Cells(j + 4,
X_col + 4)).BorderAround (xlDouble)

    output.Cells(j + 2, X_col + 6) = "S"
    output.Cells(j + 3, X_col + 6) = "G.L."
    output.Cells(j + 4, X_col + 6) = "SQe"
    output.Cells(j + 4, X_col + 6).Characters(Start:=3,
Length:=1).Font.Subscript = True
    Range(output.Cells(j + 2, X_col + 7), output.Cells(j + 4,
X_col + 7)) = Stats_final2
    Range(output.Cells(j + 2, X_col + 7), output.Cells(j + 4,
X_col + 7)).NumberFormat = "#,##0.000"
    output.Cells(j + 3, X_col + 7).NumberFormat = "#,##0"
    Range(output.Cells(j + 2, X_col + 6), output.Cells(j + 4,
X_col + 6)).Font.Bold = True

    With Range(output.Cells(j + 2, X_col + 6), output.Cells(j
+ 4, X_col + 7)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 2, X_col + 6), output.Cells(j + 4,
X_col + 7)).BorderAround (xlDouble)

  End If

  output.Cells(j + 6, 1).Value = "Seleção Y:"
  output.Cells(j + 6, 1).Font.Bold = True
  output.Cells(j + 7, 1).Value = oY.Worksheet.Name
  output.Cells(j + 8, 1).Value = oY.Address(False, False)

  output.Cells(j + 10, 1).Value = "Seleção X:"
  output.Cells(j + 10, 1).Font.Bold = True
  output.Cells(j + 11, 1).Value = oX.Worksheet.Name
  output.Cells(j + 12, 1).Value = oX.Address(False, False)

  Sheets("Out_Forward_Linear").Cells.EntireColumn.AutoFit
Sheets("Out_Forward_Linear").Cells.EntireRow.AutoFit

End Sub
```

11.6. SUB_06_LIN_STEPWISE

```
Sub Lin_Stepwise(oY As Range, oX As Range, ct As Integer,
alfa_in As Double, alfa_out As Double)
  Folha_Dados = ActiveSheet.Name

  Sheets.Add after:=Sheets(Sheets.Count)
  Sheets(ActiveSheet.Name).Name = "Out_Stepwise_Linear"
  ActiveWindow.DisplayGridlines = False

  Worksheets(Folha_Dados).Activate
  Set output = Sheets("Out_Stepwise_Linear").Range("B4")

  With Sheets("Out_Stepwise_Linear").Cells.Font
      .Name = "Times New Roman"
      .Size = 10
  End With

  Sheets("Out_Stepwise_Linear").Cells.HorizontalAlignment =
xlCenter
  Sheets("Out_Stepwise_Linear").Cells.VerticalAlignment =
xlCenter
'-------------------------------------------------------------

  Dim oX_col As Long, oX_lin As Long
  oX_col = oX.Columns.Count
  oX_lin = oX.Rows.Count - 1

  output.Cells(-1, 1).Value = "Regressão Linear: Método de
Seleção Progressiva e Regressiva (Stepwise Selection)"
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Merge
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Font.Bold = True
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Interior.Color = RGB(218, 238, 243)


  'Para o layout do quadro de p-values a escrever no excel
  'oH: Header do quadro
  'H_X: Header com nomes das variaveis a usar no index-match
  'Q_pv: vetor de p-values organizados
  ReDim oH(1 To 1, 1 To oX_col + 1), H_X(1 To 1, 1 To oX_col
+ 1)
```

```
  ReDim QF1(1 To oX_col * 3, 1 To oX_col + 1), QF2(1 To
oX_col * 3, 1 To oX_col + 1), QF3(1 To oX_col * 3, 1 To
oX_col + 1), HF(1 To oX_col * 3, 1 To oX_col + 1)
  ReDim QF4(1 To oX_col * 3, 1 To oX_col + 1), Stats1(1 To
oX_col * 3, 1 To 2), Stats2(1 To oX_col * 3, 1 To 2),
Stats3(1 To oX_col * 3, 1 To 2), pv_aux_final(1 To oX_col *
3, 1 To 1)
  ReDim col_aux(1 To oX_col * 3, 1 To oX_col)

  For col = 1 To oX_col
    oH(1, col) = oX(1, col)
    H_X(1, col) = oX(1, col)
  Next col
    oH(1, oX_col + 1) = "const"
    H_X(1, oX_col + 1) = "const"

  Range(output.Cells(1, 2), output.Cells(1, oX_col + 2)) = oH
  output.Cells(1, oX_col + 4) = "R2"
  output.Cells(1, oX_col + 4).Characters(Start:=2,
Length:=1).Font.Superscript = True
  Range(output.Cells(1, 2), output.Cells(1, oX_col +
4)).Font.Bold = True

  'Parametros para a primeira iteracao:
  'consideramos a matriz de planeamento que o utilizador
seleccionou
  aux_max = 0
  PV_in = 999
  max_PV_out = 999
  X_col = oX_col

  ReDim X(1 To oX_lin, 1 To oX_col) As Variant
  For col = 1 To oX_col
    For Lin = 1 To oX_lin
      X(Lin, col) = oX(Lin + 1, col)
    Next Lin
  Next col

  ReDim y(1 To oX_lin, 1 To 1) As Variant
  For Lin = 1 To oX_lin
    y(Lin, 1) = oY(Lin + 1, 1)
  Next Lin
  k = 1 'primeira iteracao
  ReDim R2_output(1 To 1, 1 To X_col)

  For col = 1 To X_col
    ReDim X_teste(1 To oX_lin, 1 To 1)
    For Lin = 1 To oX_lin
      X_teste(Lin, 1) = oX(Lin + 1, col)
    Next Lin

    linest_output = Application.WorksheetFunction.LinEst(y,
X_teste, ct, 1)

    R2 = Application.WorksheetFunction.Index(linest_output,
3, 1)
    R2_output(1, col) = R2
  Next col

  max_R2 = Application.WorksheetFunction.Max(R2_output)
  aux_max_r2 = Application.WorksheetFunction.Match(max_R2,
R2_output, 0)

  'Para criar a nova matriz de planeamento base, com menos
uma coluna
  'Matriz de Planeamento do modelo com a 1a variavel
seleccionada
  Call remove_col(X, aux_max_r2, X, col_X)
  Call remove_col(H_X, aux_max_r2, H_X, col_H)

  ReDim X_modelo(1 To oX_lin, 1 To 1)
  For Lin = 1 To oX_lin
    X_modelo(Lin, 1) = col_X(Lin, 1)
  Next Lin

  ReDim H_modelo(1 To 1, 1 To 2)
  H_modelo(1, 1) = col_H(1, 1)
  H_modelo(1, 2) = "const"

  For col = 1 To 2
    HF(k, col) = H_modelo(1, col)
  Next col

  'Funcao Linest
  linest_output = Application.WorksheetFunction.LinEst(y,
X_modelo, ct, 1)

  'Para pegar no output do linest e calcular os P-Values
```

```
  Call P_Value_T(1, linest_output, ct, QF, P_V, max_P_V,
aux_max, PV_in)

  'juntar os p-values no quadro a escrver na folha
  ReDim Q_pv(1 To 1, 1 To oX_col + 1)
  For col = 1 To oX_col + 1
    On Error Resume Next 'para ficar em branco quando o
index-match nao encontrar correspondencia
    Q_pv(1, col) = WorksheetFunction.Index(P_V, 1,
WorksheetFunction.Match(oH(1, col), H_modelo, 0))
  Next col

  'P-Values
  Range(output.Cells(k + 1, 2), output.Cells(k + 1, oX_col +
2)) = Q_pv
  'R2
  output.Cells(k + 1, oX_col + 4) =
Application.WorksheetFunction.Index(linest_output, 3, 1)
  Range(output.Cells(k + 1, 2), output.Cells(k + 1, oX_col +
4)).NumberFormat = "#,##0.000"

  'guardamos a informacao de cada modelo seleccionado a cada
iteracao
  For col = 1 To 2
    QF1(k, col) = QF(1, col)
    QF2(k, col) = QF(2, col)
    QF3(k, col) = QF(3, col)
    QF4(k, col) = QF(4, col)
    Stats1(k, col) =
Application.WorksheetFunction.Index(linest_output, 3, col)
    Stats2(k, col) =
Application.WorksheetFunction.Index(linest_output, 4, col)
    Stats3(k, col) =
Application.WorksheetFunction.Index(linest_output, 5, col)
  Next col

  'Auxiliar para apurar a iteracao com o modelo seleccionado:
sera o ultimo em que todas as variaveis sao significativas
  If PV_in < alfa_in Then
    pv_aux_final(k, 1) = 1
  Else
    pv_aux_final(k, 1) = 0
  End If

  'p-value in < alfa_in
  For col = 1 To oX_col
    If output.Cells(k + 1, col).Value <> "" And
output.Cells(k + 1, col).Value < alfa_in Then
      output.Cells(k + 1, col).Interior.Color = RGB(218, 238,
243)
    End If
  Next col

  'p-value in >= alfa_in
  For col = 1 To oX_col
    If output.Cells(k + 1, col).Value <> "" And
output.Cells(k + 1, col).Value >= alfa_in Then
      output.Cells(k + 1, col).Interior.Color = RGB(166, 166,
166)
    End If
  Next col

  k = k + 1

  If PV_in >= alfa_in Then 'O algoritmo para e nao temos
modelo

    Range(output.Cells(k, 2), output.Cells(k, oX_col +
2)).Interior.Color = RGB(217, 217, 217)
    output.Cells(k, oX_col + 4).Interior.Color = RGB(217,
217, 217)
    For col = 1 To oX_col + 2
      If output.Cells(k, col).Value = PV_in And PV_in >=
alfa_in Then
        output.Cells(k, col).Interior.Color = RGB(166, 166,
166)
      End If
    Next col

    With Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).BorderAround (xlDouble)
    Range(output.Cells(1, 2), output.Cells(1, oX_col +
2)).BorderAround (xlDouble)
```

```
    Range(output.Cells(1, oX_col + 4), output.Cells(k, oX_col
+ 4)).BorderAround (xlDouble)
    output.Cells(1, oX_col + 4).BorderAround (xlDouble)
    With Range(output.Cells(2, oX_col + 4), output.Cells(k,
oX_col + 4)).Borders(xlInsideHorizontal)
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With

    output.Cells(4, 2) = "ain"
    output.Cells(4, 2).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(4, 2).Characters(Start:=2,
Length:=2).Font.Subscript = True
    output.Cells(4, 2).Font.Bold = True
    output.Cells(4, 3) = alfa_in
    output.Cells(4, 3).NumberFormat = "0.000"

    output.Cells(4, 5) = "aout"
    output.Cells(4, 5).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(4, 5).Characters(Start:=2,
Length:=3).Font.Subscript = True
    output.Cells(4, 5).Font.Bold = True
    output.Cells(4, 6) = alfa_out
    output.Cells(4, 6).NumberFormat = "0.000"

    With Range(output.Cells(4, 2), output.Cells(4,
3)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(4, 2), output.Cells(4,
3)).BorderAround (xlDouble)

    With Range(output.Cells(4, 5), output.Cells(4,
6)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(4, 5), output.Cells(4,
6)).BorderAround (xlDouble)
```

```
  Else 'continuamos o algoritmo normalmente ate que a
variavel a entrar deixe de ser significativa mas todas as
restantes sejam
    Do Until PV_in >= alfa_in Or k = oX_col * 3 Or
(X_modelo_col = oX_col And max_PV_out < alfa_in And
max_PV_out < alfa_out)

      For col = 1 To UBound(X, 2)
        Call remove_col(X, col, X_temp, col_X)
        Call add_col(X_modelo, col_X, X_teste)

        linest_output =
Application.WorksheetFunction.LinEst(y, X_teste, ct, 1)

        R2 =
Application.WorksheetFunction.Index(linest_output, 3, 1)
        R2_output(1, col) = R2
      Next col

      max_R2 = Application.WorksheetFunction.Max(R2_output)
      aux_max_r2 =
Application.WorksheetFunction.Match(max_R2, R2_output, 0)

      'Matriz de Planeamento com a variavel seleccionada

      If UBound(X, 2) = 1 Then
        For Lin = 1 To oX_lin
          col_X(Lin, 1) = X(Lin, 1)
        Next Lin
        X_col = 0
      Else
        Call remove_col(X, aux_max_r2, X, col_X)
        X_col = UBound(X, 2)
      End If

      Call add_col(X_modelo, col_X, X_modelo)
      X_modelo_col = UBound(X_modelo, 2)

      Call remove_col(H_X, aux_max_r2, H_X, col_H)

      'Adicionar a coluna na penultima posicao
      ReDim H_modelo1(1 To 1, 1 To X_modelo_col + 1)
      For col = 1 To X_modelo_col - 1
        H_modelo1(1, col) = H_modelo(1, col)
```

```
      Next col
      H_modelo1(1, X_modelo_col) = col_H(1, 1)
      H_modelo1(1, X_modelo_col + 1) = "const"

      ReDim H_modelo(1 To 1, 1 To X_modelo_col + 1)
      For col = 1 To X_modelo_col + 1
        H_modelo(1, col) = H_modelo1(1, col)
      Next col

      'Funcao Linest
      linest_output = Application.WorksheetFunction.LinEst(y,
X_modelo, ct, 1)

      'Para pegar no output do linest e calcular os P-Values
      Call P_Value_T(X_modelo_col, linest_output, ct, QF,
P_V, max_P_V, aux_max, PV_in)

      'juntar os p-values no quadro a escrver na folha
      ReDim Q_pv(1 To 1, 1 To oX_col + 1)
      For col = 1 To oX_col + 1
        On Error Resume Next 'para ficar em branco quando o
index-match nao encontrar correspondencia
        Q_pv(1, col) = WorksheetFunction.Index(P_V, 1,
WorksheetFunction.Match(oH(1, col), H_modelo, 0))
      Next col

      'P-Values
      Range(output.Cells(k + 1, 2), output.Cells(k + 1,
oX_col + 2)) = Q_pv
      'R2
      output.Cells(k + 1, oX_col + 4) =
Application.WorksheetFunction.Index(linest_output, 3, 1)
      Range(output.Cells(k + 1, 2), output.Cells(k + 1,
oX_col + 4)).NumberFormat = "#,##0.000"

      For col = 1 To X_modelo_col + 1
        HF(k, col) = H_modelo(1, col)
      Next col

      PV_Col = UBound(P_V, 2) 'tem coluna para o p-value da
constante

      'p-value das restantes
      ReDim PV_out(1 To 1, 1 To PV_Col - 2)
      For col = 1 To PV_Col - 2
        PV_out(1, col) = P_V(1, col)
      Next col
      PV_out_Col = UBound(PV_out, 2)
      max_PV_out = Application.WorksheetFunction.Max(PV_out)

      If max_PV_out >= alfa_out Then 'retiramos as colunas
das variaveis nao significativas
        max_PV_out = 999
        For col = PV_out_Col To 1 Step -1
          If PV_out(1, col) >= alfa_out Then
            Call remove_col(X_modelo, col, X_modelo, col_X)
            Call add_col(X, col_X, X)
            Call remove_col(H_modelo, col, H_modelo, col_H)
            Dim H_X_col_aux As Long
            H_X_col_aux = UBound(H_X, 2)

            'Adicionar a coluna na penultima posicao
            ReDim H_X1(1 To 1, 1 To H_X_col_aux + 1)
            For col2 = 1 To H_X_col_aux - 1
              H_X1(1, col2) = H_X(1, col2)
            Next col2

            H_X1(1, H_X_col_aux) = col_H(1, 1)
            H_X1(1, H_X_col_aux + 1) = "const"

            ReDim H_X(1 To 1, 1 To H_X_col_aux + 1)
            For col2 = 1 To H_X_col_aux + 1
              H_X(1, col2) = H_X1(1, col2)
            Next col2
          Else
          End If
        Next col
      Else
      End If

      'O modelo final sera o obtido na penultima iteracao
      'guardamos a informacao de cada modelo seleccionado a
cada iteracao
      For col = 1 To X_modelo_col + 1
        QF1(k, col) = QF(1, col)
        QF2(k, col) = QF(2, col)
        QF3(k, col) = QF(3, col)
        QF4(k, col) = QF(4, col)
```

```
      Next col

      'Auxiliar para apurar a iteracao com o modelo
seleccionado: sera o ultimo em que todas as variaveis sao
significativas
      ReDim pv_temp(1 To 1, 1 To X_modelo_col - 1)
      For col = 1 To X_modelo_col - 1
        pv_temp(1, col) = QF(4, col)
      Next col
      max_pv_final =
Application.WorksheetFunction.Max(pv_temp)

      If PV_in < alfa_in And max_pv_final < alfa_out Then
        pv_aux_final(k, 1) = 1
      Else
        pv_aux_final(k, 1) = 0
      End If

      For col = 1 To 2
        Stats1(k, col) =
Application.WorksheetFunction.Index(linest_output, 3, col)
        Stats2(k, col) =
Application.WorksheetFunction.Index(linest_output, 4, col)
        Stats3(k, col) =
Application.WorksheetFunction.Index(linest_output, 5, col)
      Next col

      'p-value in > alfa_in : variavel testada nao e'
significativa
      For col = 1 To oX_col
        If output.Cells(k, col + 1).Value = "" And
output.Cells(k + 1, col + 1).Value <> "" And output.Cells(k +
1, col + 1).Value >= alfa_in Then
          output.Cells(k + 1, col + 1).Interior.Color =
RGB(166, 166, 166)
        End If
      Next col

      'p-value out > alfa_out : variaveis a sair
      For col = 1 To oX_col
        If output.Cells(k + 1, col + 1).Value > alfa_out Then
          output.Cells(k + 1, col + 1).Interior.Color =
RGB(166, 166, 166)
        End If
      Next col

      'p-value in < alfa_in : variavel testada e'
significativa
      aux = 1
      For col = 1 To oX_col
        If output.Cells(k, col + 1).Value = "" And _
          output.Cells(k + 1, col + 1).Value <> "" And
output.Cells(k + 1, col + 1).Value < alfa_in Then
          output.Cells(k + 1, col + 1).Interior.Color =
RGB(218, 238, 243)
          aux = 0
          col_aux(k + 1, 1) = col
        End If
      Next col

      For col = 1 To oX_col
        If aux = 1 And col <> col_aux(k, 1) And _
          output.Cells(k, col + 1).Value > alfa_out And _
          output.Cells(k + 1, col + 1).Value <> "" And
output.Cells(k + 1, col + 1).Value < alfa_in Then
          output.Cells(k + 1, col + 1).Interior.Color =
RGB(218, 238, 243)
          col_aux(k + 1, 1) = col
        End If
      Next col

      k = k + 1
    Loop

    With Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).BorderAround (xlDouble)
    Range(output.Cells(1, 2), output.Cells(1, oX_col +
2)).BorderAround (xlDouble)

    Range(output.Cells(1, oX_col + 4), output.Cells(k, oX_col
+ 4)).BorderAround (xlDouble)
    output.Cells(1, oX_col + 4).BorderAround (xlDouble)
```

```
    With Range(output.Cells(2, oX_col + 4), output.Cells(k,
oX_col + 4)).Borders(xlInsideHorizontal)
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With

    'Temos que determinar em que iteracao foi testado o
modelo seleccionado
    iter = 0
    Lin = UBound(pv_aux_final, 1)
    Do Until pv_aux_final(Lin, 1) = 1 Or Lin = 0
      Lin = Lin - 1
      iter = Lin
    Loop

    If iter = 0 Then
      iter = 1
    Else
    End If

    ReDim H_X(1 To 1, 1 To k + 1)
    For col = 1 To k + 1
      H_X(1, col) = HF(iter, col)
    Next col

    X_modelo_col =
Application.WorksheetFunction.Match("const", H_X, 0)

    ReDim QF_final(1 To 4, 1 To X_modelo_col)
    For col = 1 To X_modelo_col + 1
      QF_final(1, col) = QF1(iter, col)
      QF_final(2, col) = QF2(iter, col)
      QF_final(3, col) = QF3(iter, col)
      QF_final(4, col) = QF4(iter, col)
    Next col

    ReDim Stats_final1(1 To 3, 1 To 1)
    Stats_final1(1, 1) = Stats1(iter, 1)
    Stats_final1(2, 1) = Stats2(iter, 1)
    Stats_final1(3, 1) = Stats3(iter, 1)

    ReDim Stats_final2(1 To 3, 1 To 1)
    Stats_final2(1, 1) = Stats1(iter, 2)
    Stats_final2(2, 1) = Stats2(iter, 2)
    Stats_final2(3, 1) = Stats3(iter, 2)

    j = k + 3
    X_col = X_modelo_col

    output.Cells(j + 0, 1) = "Coeficientes"
    output.Cells(j + 1, 1) = "EMQ"
    output.Cells(j + 2, 1) = "s(b)"
    output.Cells(j + 2, 1).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(j + 3, 1) = "Estat. Teste"
    output.Cells(j + 4, 1) = "P-Value"
    Range(output.Cells(j + 0, 2), output.Cells(j + 0, X_col +
1)) = H_X
    Range(output.Cells(j + 0, 1), output.Cells(j + 0, X_col +
1)).Font.Bold = True
    Range(output.Cells(j + 1, 1), output.Cells(j + 4,
1)).Font.Bold = True
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
1)) = QF_final
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
1)).NumberFormat = "#,##0.000"

    With Range(output.Cells(j + 0, 1), output.Cells(j + 4,
X_col + 1)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, 1), output.Cells(j + 4, X_col +
1)).BorderAround (xlDouble)
    output.Cells(j + 0, 1).BorderAround (xlDouble)
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
1)).BorderAround (xlDouble)

    output.Cells(j + 0, X_col + 3) = "ain"
    output.Cells(j + 0, X_col + 3).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(j + 0, X_col + 3).Characters(Start:=2,
Length:=2).Font.Subscript = True
    output.Cells(j + 0, X_col + 3).Font.Bold = True
    output.Cells(j + 0, X_col + 4) = alfa_in
    output.Cells(j + 0, X_col + 4).NumberFormat = "0.000"
```

```
    With Range(output.Cells(j + 0, X_col + 3), output.Cells(j
+ 0, X_col + 4)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, X_col + 3), output.Cells(j + 0,
X_col + 4)).BorderAround (xlDouble)

    output.Cells(j + 0, X_col + 6) = "aout"
    output.Cells(j + 0, X_col + 6).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(j + 0, X_col + 6).Characters(Start:=2,
Length:=3).Font.Subscript = True
    output.Cells(j + 0, X_col + 6).Font.Bold = True
    output.Cells(j + 0, X_col + 7) = alfa_out
    output.Cells(j + 0, X_col + 7).NumberFormat = "0.000"

    With Range(output.Cells(j + 0, X_col + 6), output.Cells(j
+ 0, X_col + 7)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, X_col + 6), output.Cells(j + 0,
X_col + 7)).BorderAround (xlDouble)


    output.Cells(j + 2, X_col + 3) = "R2"
    output.Cells(j + 2, X_col + 3).Characters(Start:=2,
Length:=1).Font.Superscript = True

    output.Cells(j + 3, X_col + 3) = "Teste-F"
    output.Cells(j + 4, X_col + 3) = "SQReg"
    output.Cells(j + 4, X_col + 3).Characters(Start:=3,
Length:=3).Font.Subscript = True
    Range(output.Cells(j + 2, X_col + 4), output.Cells(j + 4,
X_col + 4)) = Stats_final1
    Range(output.Cells(j + 2, X_col + 4), output.Cells(j + 4,
X_col + 4)).NumberFormat = "#,##0.000"
    Range(output.Cells(j + 0, X_col + 3), output.Cells(j + 4,
X_col + 3)).Font.Bold = True

    With Range(output.Cells(j + 2, X_col + 3), output.Cells(j
+ 4, X_col + 4)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 2, X_col + 3), output.Cells(j + 4,
X_col + 4)).BorderAround (xlDouble)

    output.Cells(j + 2, X_col + 6) = "S"
    output.Cells(j + 3, X_col + 6) = "G.L."
    output.Cells(j + 4, X_col + 6) = "SQe"
    output.Cells(j + 4, X_col + 6).Characters(Start:=3,
Length:=1).Font.Subscript = True
    Range(output.Cells(j + 2, X_col + 7), output.Cells(j + 4,
X_col + 7)) = Stats_final2
    Range(output.Cells(j + 2, X_col + 7), output.Cells(j + 4,
X_col + 7)).NumberFormat = "#,##0.000"
    output.Cells(j + 3, X_col + 7).NumberFormat = "#,##0"
    Range(output.Cells(j + 0, X_col + 6), output.Cells(j + 4,
X_col + 6)).Font.Bold = True

    With Range(output.Cells(j + 2, X_col + 6), output.Cells(j
+ 4, X_col + 7)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 2, X_col + 6), output.Cells(j + 4,
X_col + 7)).BorderAround (xlDouble)

  End If

  output.Cells(j + 6, 1).Value = "Seleção Y:"
  output.Cells(j + 6, 1).Font.Bold = True
  output.Cells(j + 7, 1).Value = oY.Worksheet.Name
  output.Cells(j + 8, 1).Value = oY.Address(False, False)

  output.Cells(j + 10, 1).Value = "Seleção X:"
  output.Cells(j + 10, 1).Font.Bold = True
  output.Cells(j + 11, 1).Value = oX.Worksheet.Name
  output.Cells(j + 12, 1).Value = oX.Address(False, False)

  Sheets("Out_Stepwise_Linear").Cells.EntireColumn.AutoFit
  Sheets("Out_Stepwise_Linear").Cells.EntireRow.AutoFit

End Sub
```

11.7. SUB_07_LOG_BACKWARD

```
Sub Log_Backward(oY As Range, oX As Range, alfa As Double, pc
As Long)
  Folha_Dados = ActiveSheet.Name

  Sheets.Add after:=Sheets(Sheets.Count)
  Sheets(ActiveSheet.Name).Name = "Out_Backward_Logistica"
  ActiveWindow.DisplayGridlines = False

  Worksheets(Folha_Dados).Activate
  Set output = Sheets("Out_Backward_Logistica").Range("B4")

  With Sheets("Out_Backward_Logistica").Cells.Font
      .Name = "Times New Roman"
      .Size = 10
  End With

  Sheets("Out_Backward_Logistica").Cells.HorizontalAlignment
= xlCenter
  Sheets("Out_Backward_Logistica").Cells.VerticalAlignment =
xlCenter
'-----------------------------------------------------------


  Dim oX_col As Long, oX_lin As Long
  oX_col = oX.Columns.Count
  oX_lin = oX.Rows.Count - 1

  output.Cells(-1, 1).Value = "Regressão Logística: Método de
Seleção Regressiva (Backward Selection)"
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Merge
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Font.Bold = True
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Interior.Color = RGB(218, 238, 243)

  'Para o layout do quadro de p-values a escrever no excel
  'oH: Header do quadro
  'H_X: Header com nomes das variaveis a usar no index-match
  'Q_pv: vetor de p-values organizados
  ReDim oH(1 To 1, 1 To oX_col + 1), H_X(1 To 1, 1 To oX_col
+ 1), Q_pv(1 To 1, 1 To oX_col + 2)

  For col = 1 To oX_col
    oH(1, col) = oX(1, col)
    H_X(1, col) = oX(1, col)
  Next col
    oH(1, oX_col + 1) = "const"
    H_X(1, oX_col + 1) = "const"

  Range(output.Cells(1, 2), output.Cells(1, oX_col + 2)) = oH
  output.Cells(1, oX_col + 4) = "'-2*ln(L0/L1)"
  output.Cells(1, oX_col + 4).Characters(Start:=8,
Length:=1).Font.Subscript = True
  output.Cells(1, oX_col + 4).Characters(Start:=11,
Length:=1).Font.Subscript = True
  Range(output.Cells(1, 2), output.Cells(1, oX_col +
4)).Font.Bold = True

  'Parametros para a primeira iteracao:
  'consideramos a matriz de planeamento que o utilizador
seleccionou
  aux_max = 0
  max_P_V = 999
  ReDim X(1 To oX_lin, 1 To oX_col) As Variant
  For col = 1 To oX_col
    For Lin = 1 To oX_lin
      X(Lin, col) = oX(Lin + 1, col)
    Next Lin
  Next col

  ReDim y(1 To oX_lin, 1) As Variant
  For Lin = 1 To oX_lin
    y(Lin, 1) = oY(Lin + 1, 1)
  Next Lin

  k = 1 'para iniciar as iteracoes

  Do Until max_P_V < alfa Or k = oX_col + 1

    'Para criar a nova matriz de planeamento
    Call remove_col(X, aux_max, X, col_X)
    X_col = UBound(X, 2)

    'construir o header correspondente das variaveis a
mostrar no quadro
```

```
    Call remove_col(H_X, aux_max, H_X, col_H)

    'Funcao Log_est (equivalente a LINEST para regressao
logistica)
    log_est_output = Log_Est1(y, X, True)

    'Para pegar no output do LOG_EST e calcular os P-Values
    Call P_Value_N(X_col, log_est_output, 1, QF, P_V,
max_P_V, aux_max, P_V_ult)

    'juntar os p-values no quadro a escrver na folha
    ReDim Q_pv(1 To 1, 1 To oX_col + 1)
    For col = 1 To oX_col + 1
      On Error Resume Next 'para ficar em branco quando o
index-match nao encontrar correspondencia
      Q_pv(1, col) = WorksheetFunction.Index(P_V, 1,
WorksheetFunction.Match(oH(1, col), H_X, 0))
    Next col

    'P-Values
    Range(output.Cells(k + 1, 2), output.Cells(k + 1, oX_col
+ 2)) = Q_pv
    '-2xlog(Lo/L1)
    output.Cells(k + 1, oX_col + 4) =
Application.WorksheetFunction.Index(log_est_output, 4, 2)

    'Formatacao do Quadro dos P-Values
    For col = 1 To oX_col + 2
      If output.Cells(k + 1, col).Value = max_P_V And max_P_V
>= alfa Then
        output.Cells(k + 1, col).Interior.Color = RGB(217,
217, 217)
      End If
    Next col

    k = k + 1
  Loop

  Range(output.Cells(2, 2), output.Cells(k, oX_col +
4)).NumberFormat = "#,##0.000"

  With Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).Borders
      .LineStyle = xlContinuous
      .Color = RGB(166, 166, 166)
  End With
  Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).BorderAround (xlDouble)
  Range(output.Cells(1, 2), output.Cells(1, oX_col +
2)).BorderAround (xlDouble)

  Range(output.Cells(1, oX_col + 4), output.Cells(k, oX_col +
4)).BorderAround (xlDouble)
  output.Cells(1, oX_col + 4).BorderAround (xlDouble)
  With Range(output.Cells(2, oX_col + 4), output.Cells(k,
oX_col + 4)).Borders(xlInsideHorizontal)
      .LineStyle = xlContinuous
      .Color = RGB(166, 166, 166)
  End With

  j = k

  If max_P_V >= alfa Then 'Nao temos modelo

    output.Cells(k + 3, 2) = "a"
    output.Cells(k + 3, 2).Cells.Font.Name = "Symbol"
    output.Cells(k + 3, 2).Font.Bold = True
    output.Cells(k + 3, 3) = alfa
    output.Cells(k + 3, 3).NumberFormat = "0.000"

      With Range(output.Cells(k + 3, 2), output.Cells(k +
3, 3)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(k + 3, 2), output.Cells(k + 3,
3)).BorderAround (xlDouble)

  Else

  'Quadros da ultima iteracao: informacao do modelo
seleccionado pelo algoritmo

    j = k + 3
    output.Cells(j + 0, 1) = "Coeficientes"
    output.Cells(j + 1, 1) = "EMQ"
    output.Cells(j + 2, 1) = "s(b)"
```

```
    output.Cells(j + 2, 1).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(j + 3, 1) = "Estat. Teste"
    output.Cells(j + 4, 1) = "P-Value"
    Range(output.Cells(j + 0, 2), output.Cells(j + 0, X_col +
2)) = H_X
    Range(output.Cells(j + 0, 1), output.Cells(j + 0, X_col +
4)).Font.Bold = True
    Range(output.Cells(j + 1, 1), output.Cells(j + 4,
1)).Font.Bold = True
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
2)) = QF
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
2)).NumberFormat = "#,##0.000"

    With Range(output.Cells(j + 0, 1), output.Cells(j + 4,
X_col + 2)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, 1), output.Cells(j + 4, X_col +
2)).BorderAround (xlDouble)
    output.Cells(j + 0, 1).BorderAround (xlDouble)
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
2)).BorderAround (xlDouble)

    output.Cells(j + 0, X_col + 4) = "a"
    output.Cells(j + 0, X_col + 4).Cells.Font.Name = "Symbol"
    output.Cells(j + 0, X_col + 4).Font.Bold = True
    output.Cells(j + 0, X_col + 5) = alfa
    output.Cells(j + 0, X_col + 5).NumberFormat = "0.000"

    With Range(output.Cells(j + 0, X_col + 4), output.Cells(j
+ 0, X_col + 5)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, X_col + 4), output.Cells(j + 0,
X_col + 5)).BorderAround (xlDouble)


    output.Cells(j + 2, X_col + 4) =
Application.WorksheetFunction.Index(log_est_output, 3, 1)
    output.Cells(j + 2, X_col + 4).Characters(Start:=2,
Length:=1).Font.Subscript = True
    output.Cells(j + 2, X_col + 4).Characters(Start:=5,
Length:=1).Font.Subscript = True
    output.Cells(j + 2, X_col + 5) =
Application.WorksheetFunction.Index(log_est_output, 4, 1)

    output.Cells(j + 3, X_col + 4) =
Application.WorksheetFunction.Index(log_est_output, 3, 2)
    output.Cells(j + 3, X_col + 4).Characters(Start:=8,
Length:=1).Font.Subscript = True
    output.Cells(j + 3, X_col + 4).Characters(Start:=11,
Length:=1).Font.Subscript = True
    output.Cells(j + 3, X_col + 5) =
Application.WorksheetFunction.Index(log_est_output, 4, 2)

    output.Cells(j + 4, X_col + 4) =
Application.WorksheetFunction.Index(log_est_output, 5, 1)
    output.Cells(j + 4, X_col + 5) =
Application.WorksheetFunction.Index(log_est_output, 5, 2)

    With Range(output.Cells(j + 2, X_col + 4), output.Cells(j
+ 4, X_col + 5)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 2, X_col + 4), output.Cells(j + 4,
X_col + 5)).BorderAround (xlDouble)
    Range(output.Cells(j + 2, X_col + 5), output.Cells(j + 3,
X_col + 5)).NumberFormat = "#,##0.000"
    Range(output.Cells(j + 2, X_col + 4), output.Cells(j + 4,
X_col + 4)).Font.Bold = True


    Dim coef As Range
    Set coef = Range(output.Cells(j + 0, 2), output.Cells(j +
1, X_col + 2))

    Call MConfusao_ROC(oY, oX, coef, pc, Y_obs, Y_prev,
ROC_graf)

    output.Cells(j + 0, X_col + 7) = "Cut-off"
    output.Cells(j + 0, X_col + 7).Font.Bold = True
```

```
    'por defeito, colocamos o valor medio dos y previstos
    output.Cells(j + 0, X_col + 8) =
WorksheetFunction.Average(Y_prev)
    output.Cells(j + 0, X_col + 8).NumberFormat = "#,##0.000"

    With Range(output.Cells(j + 0, X_col + 7), output.Cells(j
+ 0, X_col + 8)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, X_col + 7), output.Cells(j + 0,
X_col + 8)).BorderAround (xlDouble)
    output.Cells(j + 0, X_col + 8).Interior.Color = RGB(218,
238, 243)


    'Para obter uma matriz de confusao dinamica para qualquer
ponto de corte temos que imprimir
    'os Y Observados e os Y previstos na folha

    Dim aux_col As Long
    aux_col = WorksheetFunction.Max(oX_col + 2, X_col + 11)

    output.Cells(-2, aux_col + 20).Value = "Y Obs"
    Range(output.Cells(-1, aux_col + 20), output.Cells(oX_lin
- 2, aux_col + 20)) = Y_obs

    output.Cells(-2, aux_col + 21).Value = "Prob"
    Range(output.Cells(-1, aux_col + 21), output.Cells(oX_lin
- 2, aux_col + 21)) = Y_prev

    output.Cells(-2, aux_col + 22).Value = "Cut-off"
    output.Cells(-1, aux_col + 22).FormulaR1C1 = "=R" & k + 6
& "C" & X_col + 9 & ""
    Range(output.Cells(0, aux_col + 22), output.Cells(oX_lin
- 2, aux_col + 22)).FormulaR1C1 = "=R[-1]C[0]"

    output.Cells(-2, aux_col + 23).Value = "Y Prev"
    Range(output.Cells(-1, aux_col + 23), output.Cells(oX_lin
- 2, aux_col + 23)).FormulaR1C1 = "=IF(R[0]C[-2]>R[0]C[-
1],1,0)"

    output.Cells(-2, aux_col + 24).Value = "O1 + P1"
    Range(output.Cells(-1, aux_col + 24), output.Cells(oX_lin
- 2, aux_col + 24)).FormulaR1C1 = "=if(and(R[0]C[-
4]=1,R[0]C[-1]=1),1,0)"

    output.Cells(-2, aux_col + 25).Value = "O0 + P1"
    Range(output.Cells(-1, aux_col + 25), output.Cells(oX_lin
- 2, aux_col + 25)).FormulaR1C1 = "=if(and(R[0]C[-
5]=0,R[0]C[-2]=1),1,0)"

    output.Cells(-2, aux_col + 26).Value = "O1 + P0"
    Range(output.Cells(-1, aux_col + 26), output.Cells(oX_lin
- 2, aux_col + 26)).FormulaR1C1 = "=if(and(R[0]C[-
6]=1,R[0]C[-3]=0),1,0)"

    output.Cells(-2, aux_col + 27).Value = "O0 + P0"
    Range(output.Cells(-1, aux_col + 27), output.Cells(oX_lin
- 2, aux_col + 27)).FormulaR1C1 = "=if(and(R[0]C[-
7]=0,R[0]C[-4]=0),1,0)"


    output.Cells(j + 2, X_col + 9).Value = "Observados"
    Range(output.Cells(j + 2, X_col + 9), output.Cells(j + 2,
X_col + 10)).Merge
    output.Cells(j + 3, X_col + 9).Value = "Positivos"
    output.Cells(j + 3, X_col + 10).Value = "Negativos"
    output.Cells(j + 3, X_col + 11).Value = "Total"

    output.Cells(j + 4, X_col + 7).Value = "Previstos"
    Range(output.Cells(j + 4, X_col + 7), output.Cells(j + 5,
X_col + 7)).Merge
    output.Cells(j + 4, X_col + 8).Value = "Positivos"
    output.Cells(j + 5, X_col + 8).Value = "Negativos"
    output.Cells(j + 6, X_col + 8).Value = "Total"

    'Verdadeiros Positivos: Observado=1 e Previsto=1
    output.Cells(j + 4, X_col + 9).FormulaR1C1 = "=sum(R" & 2
& "C" & aux_col + 25 & ":R" & oX_lin + 1 & "C" & aux_col + 25
& ")"

    'Falsos Negativos: Observado=1 e Previsto=0
    output.Cells(j + 4, X_col + 10).FormulaR1C1 = "=sum(R" &
2 & "C" & aux_col + 26 & ":R" & oX_lin + 1 & "C" & aux_col +
26 & ")"
```

```
    'Falsos Positivos: Observado=0 e Previsto=1
    output.Cells(j + 5, X_col + 9).FormulaR1C1 = "=sum(R" & 2
& "C" & aux_col + 27 & ":R" & oX_lin + 1 & "C" & aux_col + 27
& ")"

    'Verdadeiros Negativos= Observado=0 e Previsto=0
    output.Cells(j + 5, X_col + 10).FormulaR1C1 = "=sum(R" &
2 & "C" & aux_col + 28 & ":R" & oX_lin + 1 & "C" & aux_col +
28 & ")"

    'Totais
    Range(output.Cells(j + 6, X_col + 9), output.Cells(j + 6,
X_col + 10)) = "=sum(R[" & -2 & "]C[0]:R[" & -1 & "]C[0])"
    Range(output.Cells(j + 4, X_col + 11), output.Cells(j +
5, X_col + 11)) = "=sum(R[0]C[" & -2 & "]:R[0]C[" & -1 & "])"
    output.Cells(j + 6, X_col + 11) = "=sum(R[" & -2 & "]C["
& -2 & "]:R[" & -1 & "]C[" & -1 & "])"


    Range(output.Cells(j + 2, X_col + 9), output.Cells(j + 2,
X_col + 10)).BorderAround (xlDouble)
    Range(output.Cells(j + 2, X_col + 9), output.Cells(j + 2,
X_col + 11)).Font.Bold = True

    With Range(output.Cells(j + 3, X_col + 9), output.Cells(j
+ 3, X_col + 10)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 3, X_col + 9), output.Cells(j + 3,
X_col + 10)).BorderAround (xlDouble)
    Range(output.Cells(j + 3, X_col + 9), output.Cells(j + 3,
X_col + 10)).Font.Bold = True

    Range(output.Cells(j + 4, X_col + 7), output.Cells(j + 5,
X_col + 7)).BorderAround (xlDouble)
    Range(output.Cells(j + 4, X_col + 7), output.Cells(j + 5,
X_col + 7)).Font.Bold = True

    With Range(output.Cells(j + 4, X_col + 8), output.Cells(j
+ 5, X_col + 8)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 4, X_col + 8), output.Cells(j + 5,
X_col + 8)).BorderAround (xlDouble)
    Range(output.Cells(j + 4, X_col + 8), output.Cells(j + 6,
X_col + 8)).Font.Bold = True

    With Range(output.Cells(j + 4, X_col + 9), output.Cells(j
+ 5, X_col + 10)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 4, X_col + 9), output.Cells(j + 5,
X_col + 10)).BorderAround (xlDouble)

    Range(output.Cells(j + 2, X_col + 4), output.Cells(j + 4,
X_col + 5)).BorderAround (xlDouble)
    Range(output.Cells(j + 2, X_col + 5), output.Cells(j + 3,
X_col + 5)).NumberFormat = "#,##0.000"
    Range(output.Cells(j + 2, X_col + 4), output.Cells(j + 4,
X_col + 4)).Font.Bold = True


    Range(output.Cells(j + 6, X_col + 8), output.Cells(j + 6,
X_col + 11)).Font.Color = RGB(89, 89, 89)
    Range(output.Cells(j + 3, X_col + 11), output.Cells(j +
5, X_col + 11)).Font.Color = RGB(89, 89, 89)
    output.Cells(j + 6, X_col + 11).Font.Bold = True

    With output.Cells(j + 6, X_col + 11).Borders(xlEdgeTop)
        .LineStyle = xlDouble
        .Color = RGB(166, 166, 166)
    End With
    With output.Cells(j + 6, X_col + 11).Borders(xlEdgeLeft)
        .LineStyle = xlDouble
        .Color = RGB(166, 166, 166)
    End With

    'Print curva ROC
    '1a coluna: pontos de corte
    '2a coluna: 1-E = 1 - (VN/(VN+FP))
    '3a coluna: S = VP/(VP+FN)

    output.Cells(j + 0, X_col + 14).Value = "Dados Curva ROC"
    Range(output.Cells(j + 0, X_col + 14), output.Cells(j +
0, X_col + 16)).Merge
```

```
    output.Cells(j + 1, X_col + 14).Value = "pt corte"
    output.Cells(j + 1, X_col + 15).Value = "1-E"
    output.Cells(j + 1, X_col + 16).Value = "S"
    Range(output.Cells(j + 0, X_col + 14), output.Cells(j +
1, X_col + 16)).Font.Bold = True

    Range(output.Cells(j + 2, X_col + 14), output.Cells(j +
pc + 1, X_col + 16)) = ROC_graf
    Range(output.Cells(j + 2, X_col + 14), output.Cells(j +
pc + 1, X_col + 16)).NumberFormat = "0.00"

    Call graph_ROC(output, j, X_col, pc)
    aux_ROC = 1

  End If

  output.Cells(j + 6, 1).Value = "Seleção Y:"
  output.Cells(j + 6, 1).Font.Bold = True
  output.Cells(j + 7, 1).Value = oY.Worksheet.Name
  output.Cells(j + 8, 1).Value = oY.Address(False, False)

  output.Cells(j + 10, 1).Value = "Seleção X:"
  output.Cells(j + 10, 1).Font.Bold = True
  output.Cells(j + 11, 1).Value = oX.Worksheet.Name
  output.Cells(j + 12, 1).Value = oX.Address(False, False)

  Sheets("Out_Backward_Logistica").Cells.EntireColumn.AutoFit
  Sheets("Out_Backward_Logistica").Cells.EntireRow.AutoFit

  If aux_ROC = 1 Then
    output.Cells(-2, oX_col + 6).Value = "ATENÇÃO: recomenda-
se apagar o gráfico antes de apagar a folha!"
    output.Cells(-2, oX_col + 6).HorizontalAlignment = xlLeft
    output.Cells(-2, oX_col + 6).Characters(Start:=1,
Length:=8).Font.Bold = True
    output.Cells(-2, oX_col + 6).Characters(Start:=1,
Length:=8).Font.Color = RGB(255, 0, 0)
  End If

End Sub
```

11.8. SUB_08_LOG_FORWARD

```
Sub Log_Forward(oY As Range, oX As Range, alfa As Double, pc
As Long)
  Folha_Dados = ActiveSheet.Name

  Sheets.Add after:=Sheets(Sheets.Count)
  Sheets(ActiveSheet.Name).Name = "Out_Forward_Logistica"
  ActiveWindow.DisplayGridlines = False

  Worksheets(Folha_Dados).Activate
  Set output = Sheets("Out_Forward_Logistica").Range("B4")

  With Sheets("Out_Forward_Logistica").Cells.Font
      .Name = "Times New Roman"
      .Size = 10
  End With

  Sheets("Out_Forward_Logistica").Cells.HorizontalAlignment =
xlCenter
  Sheets("Out_Forward_Logistica").Cells.VerticalAlignment =
xlCenter
'-------------------------------------------------------------


  Dim oX_col As Long, oX_lin As Long
  oX_col = oX.Columns.Count
  oX_lin = oX.Rows.Count - 1

  output.Cells(-1, 1).Value = "Regressão Logística: Método de
Seleção Progressiva (Forward Selection)"
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Merge
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Font.Bold = True
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Interior.Color = RGB(218, 238, 243)


  'Para o layout do quadro de p-values a escrever no excel
  'oH: Header do quadro
  'H_X: Header do quadro de p-values com nomes das variaveis
a usar no index-match
  'Q_pv: vetor de p-values organizados
```

```
  ReDim oH(1 To 1, 1 To oX_col + 1), H_X(1 To 1, 1 To oX_col
+ 1)
  ReDim QF1(1 To oX_col + 2, 1 To oX_col + 1), QF2(1 To
oX_col + 2, 1 To oX_col + 1), QF3(1 To oX_col + 2, 1 To
oX_col + 1), HF(1 To oX_col + 2, 1 To oX_col + 1)
  ReDim QF4(1 To oX_col + 2, 1 To oX_col + 1), Stats1(1 To
oX_col + 2, 1 To 2), Stats2(1 To oX_col + 2, 1 To 2),
Stats3(1 To oX_col + 2, 1 To 2), pv_aux_final(1 To oX_col +
2, 1 To 1)

  For col = 1 To oX_col
    oH(1, col) = oX(1, col)
    H_X(1, col) = oX(1, col)
  Next col
    oH(1, oX_col + 1) = "const"
    H_X(1, oX_col + 1) = "const"

  Range(output.Cells(1, 2), output.Cells(1, oX_col + 2)) = oH
  output.Cells(1, oX_col + 4) = "'-2*ln(L0/L1)"
  output.Cells(1, oX_col + 4).Characters(Start:=8,
Length:=1).Font.Subscript = True
  output.Cells(1, oX_col + 4).Characters(Start:=11,
Length:=1).Font.Subscript = True
  Range(output.Cells(1, 2), output.Cells(1, oX_col +
4)).Font.Bold = True

  'Parametros para a primeira iteracao:
  'consideramos a matriz de planeamento que o utilizador
seleccionou
  aux_max = 0
  PV_in = 999
  X_col = oX_col

  ReDim X(1 To oX_lin, 1 To oX_col) As Variant
  For col = 1 To oX_col
    For Lin = 1 To oX_lin
      X(Lin, col) = oX(Lin + 1, col)
    Next Lin
  Next col

  ReDim y(1 To oX_lin, 1) As Variant
  For Lin = 1 To oX_lin
    y(Lin, 1) = oY(Lin + 1, 1)
  Next Lin
```

```
  k = 1 'primeira iteracao
  ReDim Rz_Ver_output(1 To 1, 1 To X_col)
  For col = 1 To X_col
    ReDim X_teste(1 To oX_lin, 1 To 1)
    For Lin = 1 To oX_lin
      X_teste(Lin, 1) = oX(Lin + 1, col)
    Next Lin

    log_est_output = Log_Est1(y, X_teste, False)

    Rz_Ver =
Application.WorksheetFunction.Index(log_est_output, 1, 1)
    Rz_Ver_output(1, col) = Rz_Ver
  Next col

  min_Rz_Ver =
Application.WorksheetFunction.Min(Rz_Ver_output)
  aux_min_Rz_Ver =
Application.WorksheetFunction.Match(min_Rz_Ver,
Rz_Ver_output, 0)

  'Para criar a nova matriz de planeamento base, com menos
uma coluna
  'Matriz de Planeamento do modelo com a 1a variavel
seleccionada
  Call remove_col(X, aux_min_Rz_Ver, X, col_X)
  Call remove_col(H_X, aux_min_Rz_Ver, H_X, col_H)

  ReDim X_modelo(1 To oX_lin, 1 To 1)
  For Lin = 1 To oX_lin
    X_modelo(Lin, 1) = col_X(Lin, 1)
  Next Lin
  ReDim H_modelo(1 To 1, 1 To 2)
  H_modelo(1, 1) = col_H(1, 1)
  H_modelo(1, 2) = "const"

  For col = 1 To 2
    HF(k, col) = H_modelo(1, col)
  Next col

  'Funcao Log_est (equivalente a LINEST para regressao
logistica)
  log_est_output = Log_Est1(y, X_modelo, True)
```

```
  'Para pegar no output do Log_est e calcular os P-Values
  Call P_Value_N(1, log_est_output, 1, QF, P_V, max_P_V,
aux_max, PV_in)

  'juntar os p-values no quadro a escrver na folha
  ReDim Q_pv(1 To 1, 1 To oX_col + 1)
  For col = 1 To oX_col + 1
    On Error Resume Next 'para ficar em branco quando o
index-match nao encontrar correspondencia
    Q_pv(1, col) = WorksheetFunction.Index(P_V, 1,
WorksheetFunction.Match(oH(1, col), H_modelo, 0))
  Next col

  'P-Values
  Range(output.Cells(k + 1, 2), output.Cells(k + 1, oX_col +
2)) = Q_pv
  '-2xlog(Lo/L1)
  output.Cells(k + 1, oX_col + 4) =
Application.WorksheetFunction.Index(log_est_output, 4, 2)
  Range(output.Cells(k + 1, 2), output.Cells(k + 1, oX_col +
4)).NumberFormat = "#,##0.000"

  'O modelo final sera o obtido na penultima iteracao
  'informacao do modelo seleccionado a cada iteracao
  For col = 1 To 2
    QF1(k, col) = QF(1, col)
    QF2(k, col) = QF(2, col)
    QF3(k, col) = QF(3, col)
    QF4(k, col) = QF(4, col)
  Next col

  Stats1(k, 1) =
Application.WorksheetFunction.Index(log_est_output, 4, 1)
  Stats2(k, 1) =
Application.WorksheetFunction.Index(log_est_output, 4, 2)
  Stats3(k, 1) =
Application.WorksheetFunction.Index(log_est_output, 5, 2)

  'Auxiliar para apurar a iteracao com o modelo seleccionado:
sera o ultimo em que todas as variaveis sao significativas
  If PV_in < alfa Then
    pv_aux_final(k, 1) = 1
  Else
    pv_aux_final(k, 1) = 0
  End If

  k = k + 1

  If PV_in >= alfa Then 'O algoritmo para e nao temos modelo

    Range(output.Cells(k, 2), output.Cells(k, oX_col +
2)).Interior.Color = RGB(217, 217, 217)
    output.Cells(k, oX_col + 4).Interior.Color = RGB(217,
217, 217)
    For col = 1 To oX_col + 2
      If output.Cells(k, col).Value = PV_in And PV_in >= alfa
Then
        output.Cells(k, col).Interior.Color = RGB(166, 166,
166)
      End If
    Next col

    With Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).BorderAround (xlDouble)
    Range(output.Cells(1, 2), output.Cells(1, oX_col +
2)).BorderAround (xlDouble)

    Range(output.Cells(1, oX_col + 4), output.Cells(k, oX_col
+ 4)).BorderAround (xlDouble)
    output.Cells(1, oX_col + 4).BorderAround (xlDouble)
    With Range(output.Cells(2, oX_col + 4), output.Cells(k,
oX_col + 4)).Borders(xlInsideHorizontal)
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With

    output.Cells(4, 2) = "a"
    output.Cells(4, 2).Cells.Font.Name = "Symbol"
    output.Cells(4, 2).Font.Bold = True
    output.Cells(4, 3) = alfa
    output.Cells(4, 3).NumberFormat = "0.000"
```

```
    With Range(output.Cells(4, 2), output.Cells(4,
3)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(4, 2), output.Cells(4,
3)).BorderAround (xlDouble)

  Else 'continuamos o algoritmo normalmente
    Do Until PV_in >= alfa Or k = oX_col + 1

      For col = 1 To X_col
        Call remove_col(X, col, X_temp, col_X)
        Call add_col(X_modelo, col_X, X_teste)

        log_est_output = Log_Est1(y, X_teste, False)

        Rz_Ver =
Application.WorksheetFunction.Index(log_est_output, 1, 1)
        Rz_Ver_output(1, col) = Rz_Ver
      Next col

      min_Rz_Ver =
Application.WorksheetFunction.Min(Rz_Ver_output)
      aux_min_Rz_Ver =
Application.WorksheetFunction.Match(min_Rz_Ver,
Rz_Ver_output, 0)

      'Matriz de planeamento base, com menos uma coluna
      'Matriz de Planeamento com a variavel seleccionada
      If UBound(X, 2) = 1 Then
        For Lin = 1 To oX_lin
          col_X(Lin, 1) = X(Lin, 1)
        Next Lin
        X_col = 0
      Else
        Call remove_col(X, aux_min_Rz_Ver, X, col_X)
        X_col = UBound(X, 2)
      End If

      Call add_col(X_modelo, col_X, X_modelo)
      X_col = UBound(X, 2)
      X_modelo_col = UBound(X_modelo, 2)

      Call remove_col(H_X, aux_min_Rz_Ver, H_X, col_H)

      'Adicionar a coluna na penultima posicao
      ReDim H_modelo1(1 To 1, 1 To k + 1)
      For col = 1 To k - 1
        H_modelo1(1, col) = H_modelo(1, col)
      Next col
        H_modelo1(1, k) = col_H(1, 1)
        H_modelo1(1, k + 1) = "const"

      ReDim H_modelo(1 To 1, 1 To k + 1)
      For col = 1 To k + 1
        H_modelo(1, col) = H_modelo1(1, col)
      Next col

      'Funcao Log_est (equivalente a LINEST para regressao
logistica)
      log_est_output = Log_Est1(y, X_modelo, True)

      'Para pegar no output do linest e calcular os P-Values
      Call P_Value_N(X_modelo_col, log_est_output, 1, QF,
P_V, max_P_V, aux_max, PV_in)

      'juntar os p-values no quadro a escrver na folha
      ReDim Q_pv(1 To 1, 1 To oX_col + 1)
      For col = 1 To oX_col + 1
        On Error Resume Next 'para ficar em branco quando o
index-match nao encontrar correspondencia
        Q_pv(1, col) = WorksheetFunction.Index(P_V, 1,
WorksheetFunction.Match(oH(1, col), H_modelo, 0))
      Next col

      'P-Values
      Range(output.Cells(k + 1, 2), output.Cells(k + 1,
oX_col + 2)) = Q_pv
      '-2xlog(Lo/L1)
      output.Cells(k + 1, oX_col + 4) =
Application.WorksheetFunction.Index(log_est_output, 4, 2)
      Range(output.Cells(k + 1, 2), output.Cells(k + 1,
oX_col + 4)).NumberFormat = "#,##0.000"

      'O modelo final sera o obtido na penultima iteracao
      'informacao do modelo seleccionado a cada iteracao
      For col = 1 To X_modelo_col + 1
```

```
        HF(k, col) = H_modelo(1, col)
      Next col

      For col = 1 To X_modelo_col + 1
        QF1(k, col) = QF(1, col)
        QF2(k, col) = QF(2, col)
        QF3(k, col) = QF(3, col)
        QF4(k, col) = QF(4, col)
      Next col

      Stats1(k, 1) =
Application.WorksheetFunction.Index(log_est_output, 4, 1)
      Stats2(k, 1) =
Application.WorksheetFunction.Index(log_est_output, 4, 2)
      Stats3(k, 1) =
Application.WorksheetFunction.Index(log_est_output, 5, 2)


      'Auxiliar para apurar a iteracao com o modelo
seleccionado: sera o ultimo em que todas as variaveis sao
significativas
      ReDim pv_temp(1 To 1, 1 To X_modelo_col)
      For col = 1 To X_modelo_col
        pv_temp(1, col) = QF(4, col)
      Next col
      max_pv_final =
Application.WorksheetFunction.Max(pv_temp)

      If max_pv_final < alfa Then
        pv_aux_final(k, 1) = 1
      Else
        pv_aux_final(k, 1) = 0
      End If

      k = k + 1
    Loop

    Range(output.Cells(k, 2), output.Cells(k, oX_col +
2)).Interior.Color = RGB(217, 217, 217)
    output.Cells(k, oX_col + 4).Interior.Color = RGB(217,
217, 217)
    For col = 1 To oX_col + 2
      If output.Cells(k, col).Value = PV_in And PV_in >= alfa
Then
        output.Cells(k, col).Interior.Color = RGB(166, 166,
166)
      End If
    Next col

    With Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).BorderAround (xlDouble)
    Range(output.Cells(1, 2), output.Cells(1, oX_col +
2)).BorderAround (xlDouble)

    Range(output.Cells(1, oX_col + 4), output.Cells(k, oX_col
+ 4)).BorderAround (xlDouble)
    output.Cells(1, oX_col + 4).BorderAround (xlDouble)
    With Range(output.Cells(2, oX_col + 4), output.Cells(k,
oX_col + 4)).Borders(xlInsideHorizontal)
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With


    'Temos que determinar em que iteracao foi testado o
modelo seleccionado
    iter = 0
    Lin = UBound(pv_aux_final, 1)
    Do Until pv_aux_final(Lin, 1) = 1
      Lin = Lin - 1
      iter = Lin
    Loop

    ReDim H_X(1 To 1, 1 To k + 1)
    For col = 1 To k + 1
      H_X(1, col) = HF(iter, col)
    Next col

    X_modelo_col =
Application.WorksheetFunction.Match("const", H_X, 0)

    ReDim QF_final(1 To 4, 1 To X_modelo_col + 2)
    For col = 1 To X_modelo_col + 2
```

```
      QF_final(1, col) = QF1(iter, col)
      QF_final(2, col) = QF2(iter, col)
      QF_final(3, col) = QF3(iter, col)
      QF_final(4, col) = QF4(iter, col)
    Next col

    ReDim Stats_final(1 To 3, 1 To 1)
    Stats_final(1, 1) = Stats1(iter, 1)
    Stats_final(2, 1) = Stats2(iter, 1)
    Stats_final(3, 1) = Stats3(iter, 1)

    j = k + 3
    X_col = X_modelo_col

    output.Cells(j + 0, 1) = "Coeficientes"
    output.Cells(j + 1, 1) = "EMQ"
    output.Cells(j + 2, 1) = "s(b)"
    output.Cells(j + 2, 1).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(j + 3, 1) = "Estat. Teste"
    output.Cells(j + 4, 1) = "P-Value"
    Range(output.Cells(j + 0, 2), output.Cells(j + 0, X_col +
1)) = H_X
    Range(output.Cells(j + 0, 1), output.Cells(j + 0, X_col +
2)).Font.Bold = True
    Range(output.Cells(j + 1, 1), output.Cells(j + 4,
1)).Font.Bold = True
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
1)) = QF_final
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
1)).NumberFormat = "#,##0.000"

    With Range(output.Cells(j + 0, 1), output.Cells(j + 4,
X_col + 1)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, 1), output.Cells(j + 4, X_col +
1)).BorderAround (xlDouble)
    output.Cells(j + 0, 1).BorderAround (xlDouble)
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
1)).BorderAround (xlDouble)

    output.Cells(j + 0, X_col + 3) = "a"
    output.Cells(j + 0, X_col + 3).Cells.Font.Name = "Symbol"
    output.Cells(j + 0, X_col + 3).Font.Bold = True
    output.Cells(j + 0, X_col + 4) = alfa
    output.Cells(j + 0, X_col + 4).NumberFormat = "0.000"

    With Range(output.Cells(j + 0, X_col + 3), output.Cells(j
+ 0, X_col + 4)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, X_col + 3), output.Cells(j + 0,
X_col + 4)).BorderAround (xlDouble)


    output.Cells(j + 2, X_col + 3) =
Application.WorksheetFunction.Index(log_est_output, 3, 1)
    output.Cells(j + 2, X_col + 3).Characters(Start:=2,
Length:=1).Font.Subscript = True
    output.Cells(j + 2, X_col + 3).Characters(Start:=5,
Length:=1).Font.Subscript = True
    output.Cells(j + 2, X_col + 4) = Stats_final(1, 1)

    output.Cells(j + 3, X_col + 3) =
Application.WorksheetFunction.Index(log_est_output, 3, 2)
    output.Cells(j + 3, X_col + 3).Characters(Start:=8,
Length:=1).Font.Subscript = True
    output.Cells(j + 3, X_col + 3).Characters(Start:=11,
Length:=1).Font.Subscript = True
    output.Cells(j + 3, X_col + 4) = Stats_final(2, 1)

    output.Cells(j + 4, X_col + 3) =
Application.WorksheetFunction.Index(log_est_output, 5, 1)
    output.Cells(j + 4, X_col + 4) = Stats_final(3, 1)

    With Range(output.Cells(j + 2, X_col + 3), output.Cells(j
+ 4, X_col + 4)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 2, X_col + 3), output.Cells(j + 4,
X_col + 4)).BorderAround (xlDouble)
    Range(output.Cells(j + 2, X_col + 4), output.Cells(j + 3,
X_col + 4)).NumberFormat = "#,##0.000"
```

```
    Range(output.Cells(j + 2, X_col + 3), output.Cells(j + 4,
X_col + 3)).Font.Bold = True

    Dim coef As Range
    Set coef = Range(output.Cells(j + 0, 2), output.Cells(j +
1, X_col + 1))

    Call MConfusao_ROC(oY, oX, coef, pc, Y_obs, Y_prev,
ROC_graf)

    output.Cells(j + 0, X_col + 6) = "Cut-off"
    output.Cells(j + 0, X_col + 6).Font.Bold = True
    'por defeito, colocamos o valor medio dos y previstos
    output.Cells(j + 0, X_col + 7) =
WorksheetFunction.Average(Y_prev)
    output.Cells(j + 0, X_col + 7).NumberFormat = "#,##0.000"

    With Range(output.Cells(j + 0, X_col + 6), output.Cells(j
+ 0, X_col + 7)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, X_col + 6), output.Cells(j + 0,
X_col + 7)).BorderAround (xlDouble)
    output.Cells(j + 0, X_col + 7).Interior.Color = RGB(218,
238, 243)

    'Para obter uma matriz de confusao dinamica para qualquer
ponto de corte temos que imprimir
    'os Y Observados e os Y previstos na folha

    Dim aux_col As Long
    aux_col = WorksheetFunction.Max(oX_col + 2, X_col + 11)

    output.Cells(-2, aux_col + 20).Value = "Y Obs"
    Range(output.Cells(-1, aux_col + 20), output.Cells(oX_lin
- 2, aux_col + 20)) = Y_obs

    output.Cells(-2, aux_col + 21).Value = "Prob"
    Range(output.Cells(-1, aux_col + 21), output.Cells(oX_lin
- 2, aux_col + 21)) = Y_prev

    output.Cells(-2, aux_col + 22).Value = "Cut-off"
    output.Cells(-1, aux_col + 22).FormulaR1C1 = "=R" & k + 6
& "C" & X_col + 8 & ""
    Range(output.Cells(0, aux_col + 22), output.Cells(oX_lin
- 2, aux_col + 22)).FormulaR1C1 = "=R[-1]C[0]"

    output.Cells(-2, aux_col + 23).Value = "Y Prev"
    Range(output.Cells(-1, aux_col + 23), output.Cells(oX_lin
- 2, aux_col + 23)).FormulaR1C1 = "=IF(R[0]C[-2]>R[0]C[-
1],1,0)"

    output.Cells(-2, aux_col + 24).Value = "O1 + P1"
    Range(output.Cells(-1, aux_col + 24), output.Cells(oX_lin
- 2, aux_col + 24)).FormulaR1C1 = "=if(and(R[0]C[-
4]=1,R[0]C[-1]=1),1,0)"

    output.Cells(-2, aux_col + 25).Value = "O0 + P1"
    Range(output.Cells(-1, aux_col + 25), output.Cells(oX_lin
- 2, aux_col + 25)).FormulaR1C1 = "=if(and(R[0]C[-
5]=0,R[0]C[-2]=1),1,0)"

    output.Cells(-2, aux_col + 26).Value = "O1 + P0"
    Range(output.Cells(-1, aux_col + 26), output.Cells(oX_lin
- 2, aux_col + 26)).FormulaR1C1 = "=if(and(R[0]C[-
6]=1,R[0]C[-3]=0),1,0)"

    output.Cells(-2, aux_col + 27).Value = "O0 + P0"
    Range(output.Cells(-1, aux_col + 27), output.Cells(oX_lin
- 2, aux_col + 27)).FormulaR1C1 = "=if(and(R[0]C[-
7]=0,R[0]C[-4]=0),1,0)"

    output.Cells(j + 2, X_col + 8).Value = "Observados"
    Range(output.Cells(j + 2, X_col + 8), output.Cells(j + 2,
X_col + 9)).Merge
    output.Cells(j + 3, X_col + 8).Value = "Positivos"
    output.Cells(j + 3, X_col + 9).Value = "Negativos"
    output.Cells(j + 3, X_col + 10).Value = "Total"

    output.Cells(j + 4, X_col + 6).Value = "Previstos"
    Range(output.Cells(j + 4, X_col + 6), output.Cells(j + 5,
X_col + 6)).Merge
    output.Cells(j + 4, X_col + 7).Value = "Positivos"
    output.Cells(j + 5, X_col + 7).Value = "Negativos"
    output.Cells(j + 6, X_col + 7).Value = "Total"
```

```
    'Verdadeiros Positivos: Observado=1 e Previsto=1
    output.Cells(j + 4, X_col + 8).FormulaR1C1 = "=sum(R" & 2
& "C" & aux_col + 25 & ":R" & oX_lin + 1 & "C" & aux_col + 25
& ")"

    'Falsos Negativos: Observado=1 e Previsto=0
    output.Cells(j + 4, X_col + 9).FormulaR1C1 = "=sum(R" & 2
& "C" & aux_col + 26 & ":R" & oX_lin + 1 & "C" & aux_col + 26
& ")"

    'Falsos Positivos: Observado=0 e Previsto=1
    output.Cells(j + 5, X_col + 8).FormulaR1C1 = "=sum(R" & 2
& "C" & aux_col + 27 & ":R" & oX_lin + 1 & "C" & aux_col + 27
& ")"

    'Verdadeiros Negativos= Observado=0 e Previsto=0
    output.Cells(j + 5, X_col + 9).FormulaR1C1 = "=sum(R" & 2
& "C" & aux_col + 28 & ":R" & oX_lin + 1 & "C" & aux_col + 28
& ")"

    'Totais
    Range(output.Cells(j + 6, X_col + 8), output.Cells(j + 6,
X_col + 9)) = "=sum(R[" & -2 & "]C[0]:R[" & -1 & "]C[0])"
    Range(output.Cells(j + 4, X_col + 10), output.Cells(j +
5, X_col + 10)) = "=sum(R[0]C[" & -2 & "]:R[0]C[" & -1 & "])"
    output.Cells(j + 6, X_col + 10) = "=sum(R[" & -2 & "]C["
& -2 & "]:R[" & -1 & "]C[" & -1 & "])"


    Range(output.Cells(j + 2, X_col + 8), output.Cells(j + 2,
X_col + 9)).BorderAround (xlDouble)
    Range(output.Cells(j + 2, X_col + 8), output.Cells(j + 2,
X_col + 10)).Font.Bold = True

    With Range(output.Cells(j + 3, X_col + 8), output.Cells(j
+ 3, X_col + 9)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 3, X_col + 8), output.Cells(j + 3,
X_col + 9)).BorderAround (xlDouble)
    Range(output.Cells(j + 3, X_col + 8), output.Cells(j + 3,
X_col + 9)).Font.Bold = True
    Range(output.Cells(j + 4, X_col + 6), output.Cells(j + 5,
X_col + 6)).BorderAround (xlDouble)
    Range(output.Cells(j + 4, X_col + 6), output.Cells(j + 5,
X_col + 6)).Font.Bold = True

    With Range(output.Cells(j + 4, X_col + 7), output.Cells(j
+ 5, X_col + 7)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 4, X_col + 7), output.Cells(j + 5,
X_col + 7)).BorderAround (xlDouble)
    Range(output.Cells(j + 4, X_col + 7), output.Cells(j + 6,
X_col + 7)).Font.Bold = True

    With Range(output.Cells(j + 4, X_col + 8), output.Cells(j
+ 5, X_col + 9)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 4, X_col + 8), output.Cells(j + 5,
X_col + 9)).BorderAround (xlDouble)

    Range(output.Cells(j + 6, X_col + 7), output.Cells(j + 6,
X_col + 10)).Font.Color = RGB(89, 89, 89)
    Range(output.Cells(j + 3, X_col + 10), output.Cells(j +
5, X_col + 10)).Font.Color = RGB(89, 89, 89)
    output.Cells(j + 6, X_col + 10).Font.Bold = True

    With output.Cells(j + 6, X_col + 10).Borders(xlEdgeTop)
        .LineStyle = xlDouble
        .Color = RGB(166, 166, 166)
    End With
    With output.Cells(j + 6, X_col + 10).Borders(xlEdgeLeft)
        .LineStyle = xlDouble
        .Color = RGB(166, 166, 166)
    End With

    'Print curva ROC
    '1a coluna: pontos de corte
    '2a coluna: 1-E = 1 - (VN/(VN+FP))
    '3a coluna: S = VP/(VP+FN)

    output.Cells(j + 0, X_col + 14).Value = "Dados Curva ROC"
```

```
    Range(output.Cells(j + 0, X_col + 14), output.Cells(j +
0, X_col + 16)).Merge
    output.Cells(j + 1, X_col + 14).Value = "pt corte"
    output.Cells(j + 1, X_col + 15).Value = "1-E"
    output.Cells(j + 1, X_col + 16).Value = "S"
    Range(output.Cells(j + 0, X_col + 14), output.Cells(j +
1, X_col + 16)).Font.Bold = True

    Range(output.Cells(j + 2, X_col + 14), output.Cells(j +
pc + 1, X_col + 16)) = ROC_graf
    Range(output.Cells(j + 2, X_col + 14), output.Cells(j +
pc + 1, X_col + 16)).NumberFormat = "0.00"

    Call graph_ROC(output, j, X_col, pc)
    aux_ROC = 1

  End If

  output.Cells(j + 6, 1).Value = "Seleção Y:"
  output.Cells(j + 6, 1).Font.Bold = True
  output.Cells(j + 7, 1).Value = oY.Worksheet.Name
  output.Cells(j + 8, 1).Value = oY.Address(False, False)

  output.Cells(j + 10, 1).Value = "Seleção X:"
  output.Cells(j + 10, 1).Font.Bold = True
  output.Cells(j + 11, 1).Value = oX.Worksheet.Name
  output.Cells(j + 12, 1).Value = oX.Address(False, False)

  Sheets("Out_Forward_Logistica").Cells.EntireColumn.AutoFit
  Sheets("Out_Forward_Logistica").Cells.EntireRow.AutoFit

  If aux_ROC = 1 Then
    output.Cells(-2, oX_col + 6).Value = "ATENÇÃO: recomenda-
se apagar o gráfico antes de apagar a folha!"
    output.Cells(-2, oX_col + 6).HorizontalAlignment = xlLeft
    output.Cells(-2, oX_col + 6).Characters(Start:=1,
Length:=8).Font.Bold = True
    output.Cells(-2, oX_col + 6).Characters(Start:=1,
Length:=8).Font.Color = RGB(255, 0, 0)
  End If


End Sub
```

11.9. SUB_09_LOG_STEPWISE

```
Sub Log_Stepwise(oY As Range, oX As Range, alfa_in As Double,
alfa_out As Double, pc As Long)
  Folha_Dados = ActiveSheet.Name

  Sheets.Add after:=Sheets(Sheets.Count)
  Sheets(ActiveSheet.Name).Name = "Out_Stepwise_Logistica"
  ActiveWindow.DisplayGridlines = False

  Worksheets(Folha_Dados).Activate
  Set output = Sheets("Out_Stepwise_Logistica").Range("B4")

  With Sheets("Out_Stepwise_Logistica").Cells.Font
      .Name = "Times New Roman"
      .Size = 10
  End With

  Sheets("Out_Stepwise_Logistica").Cells.HorizontalAlignment
= xlCenter
  Sheets("Out_Stepwise_Logistica").Cells.VerticalAlignment =
xlCenter
'-----------------------------------------------------------
  Dim oX_col As Long, oX_lin As Long
  oX_col = oX.Columns.Count
  oX_lin = oX.Rows.Count - 1

  output.Cells(-1, 1).Value = "Regressão Logística: Método de
Seleção Progressiva e Regressiva (Stepwise Selection)"
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Merge
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Font.Bold = True
  Range(output.Cells(-1, 1), output.Cells(-1, oX_col +
4)).Interior.Color = RGB(218, 238, 243)


  'Para o layout do quadro de p-values a escrever no excel
  'oH: Header do quadro
  'H_X: Header com nomes das variaveis a usar no index-match
  'Q_pv: vetor de p-values organizados
  ReDim oH(1 To 1, 1 To oX_col + 1), H_X(1 To 1, 1 To oX_col
+ 1)
```

```
  ReDim QF1(1 To oX_col * 3, 1 To oX_col + 1), QF2(1 To
oX_col * 3, 1 To oX_col + 1), QF3(1 To oX_col * 3, 1 To
oX_col + 1), HF(1 To oX_col * 3, 1 To oX_col + 1)
  ReDim QF4(1 To oX_col * 3, 1 To oX_col + 1), Stats1(1 To
oX_col * 3, 1 To 2), Stats2(1 To oX_col * 3, 1 To 2),
Stats3(1 To oX_col * 3, 1 To 2), pv_aux_final(1 To oX_col *
3, 1 To 1)
  ReDim col_aux(1 To oX_col * 3, 1 To oX_col)

  For col = 1 To oX_col
    oH(1, col) = oX(1, col)
    H_X(1, col) = oX(1, col)
  Next col
    oH(1, oX_col + 1) = "const"
    H_X(1, oX_col + 1) = "const"

  Range(output.Cells(1, 2), output.Cells(1, oX_col + 2)) = oH
  output.Cells(1, oX_col + 4) = "'-2*ln(L0/L1)"
  output.Cells(1, oX_col + 4).Characters(Start:=8,
Length:=1).Font.Subscript = True
  output.Cells(1, oX_col + 4).Characters(Start:=11,
Length:=1).Font.Subscript = True
  Range(output.Cells(1, 2), output.Cells(1, oX_col +
4)).Font.Bold = True



  'Parametros para a primeira iteracao:
  'consideramos a matriz de planeamento que o utilizador
seleccionou
  '  Dim PV_in As Double, max_PV_out As Double
  aux_max = 0
  PV_in = 999
  max_PV_out = 999
  X_col = oX_col

  ReDim X(1 To oX_lin, 1 To oX_col) As Variant
  For col = 1 To oX_col
    For Lin = 1 To oX_lin
      X(Lin, col) = oX(Lin + 1, col)
    Next Lin
  Next col

  ReDim y(1 To oX_lin, 1 To 1) As Variant
  For Lin = 1 To oX_lin
    y(Lin, 1) = oY(Lin + 1, 1)
  Next Lin

  k = 1 'primeira iteracao
  ReDim Rz_Ver_output(1 To 1, 1 To X_col)

  For col = 1 To X_col
    ReDim X_teste(1 To oX_lin, 1 To 1)
    For Lin = 1 To oX_lin
      X_teste(Lin, 1) = oX(Lin + 1, col)
    Next Lin

    log_est_output = Log_Est1(y, X_teste, False)

    Rz_Ver =
Application.WorksheetFunction.Index(log_est_output, 1, 1)
    Rz_Ver_output(1, col) = Rz_Ver
  Next col

  min_Rz_Ver =
Application.WorksheetFunction.Min(Rz_Ver_output)
  aux_min_Rz_Ver =
Application.WorksheetFunction.Match(min_Rz_Ver,
Rz_Ver_output, 0)

  'Para criar a nova matriz de planeamento base, com menos
uma coluna
  'Matriz de Planeamento do modelo com a 1a variavel
seleccionada
  Call remove_col(X, aux_min_Rz_Ver, X, col_X)
  Call remove_col(H_X, aux_min_Rz_Ver, H_X, col_H)

  ReDim X_modelo(1 To oX_lin, 1 To 1)
  For Lin = 1 To oX_lin
    X_modelo(Lin, 1) = col_X(Lin, 1)
  Next Lin

  ReDim H_modelo(1 To 1, 1 To 2)
  H_modelo(1, 1) = col_H(1, 1)
  H_modelo(1, 2) = "const"

  For col = 1 To 2
    HF(k, col) = H_modelo(1, col)
```

```
  Next col

  'Funcao Log_est (equivalente a LINEST para regressao
logistica)
  log_est_output = Log_Est1(y, X_modelo, True)

  'Para pegar no output do Log_est e calcular os P-Values
  Call P_Value_N(1, log_est_output, 1, QF, P_V, max_P_V,
aux_max, PV_in)

  'juntar os p-values no quadro a escrver na folha
  ReDim Q_pv(1 To 1, 1 To oX_col + 1)
  For col = 1 To oX_col + 1
    On Error Resume Next 'para ficar em branco quando o
index-match nao encontrar correspondencia
    Q_pv(1, col) = WorksheetFunction.Index(P_V, 1,
WorksheetFunction.Match(oH(1, col), H_modelo, 0))
  Next col

  'P-Values
  Range(output.Cells(k + 1, 2), output.Cells(k + 1, oX_col +
2)) = Q_pv
  '-2xlog(Lo/L1)
  output.Cells(k + 1, oX_col + 4) =
Application.WorksheetFunction.Index(log_est_output, 4, 2)
  Range(output.Cells(k + 1, 2), output.Cells(k + 1, oX_col +
4)).NumberFormat = "#,##0.000"


  'guardamos a informacao de cada modelo seleccionado a cada
iteracao
  For col = 1 To 2
    QF1(k, col) = QF(1, col)
    QF2(k, col) = QF(2, col)
    QF3(k, col) = QF(3, col)
    QF4(k, col) = QF(4, col)
  Next col

  'Auxiliar para apurar a iteracao com o modelo seleccionado:
sera o ultimo em que todas as variaveis sao significativas
  If PV_in < alfa_in Then
    pv_aux_final(k, 1) = 1
  Else
    pv_aux_final(k, 1) = 0
  End If

  Stats1(k, col) =
Application.WorksheetFunction.Index(log_est_output, 4, 1)
  Stats2(k, col) =
Application.WorksheetFunction.Index(log_est_output, 4, 2)
  Stats3(k, col) =
Application.WorksheetFunction.Index(log_est_output, 5, 2)

  'p-value in < alfa_in
  For col = 1 To oX_col
    If output.Cells(k + 1, col).Value <> "" And
output.Cells(k + 1, col).Value < alfa_in Then
      output.Cells(k + 1, col).Interior.Color = RGB(218, 238,
243)
    End If
  Next col

  'p-value in >= alfa_in
  For col = 1 To oX_col
    If output.Cells(k + 1, col).Value <> "" And
output.Cells(k + 1, col).Value >= alfa_in Then
      output.Cells(k + 1, col).Interior.Color = RGB(166, 166,
166)
    End If
  Next col

  k = k + 1

  If PV_in >= alfa_in Then 'O algoritmo para e nao temos
modelo

    Range(output.Cells(k, 2), output.Cells(k, oX_col +
2)).Interior.Color = RGB(217, 217, 217)
    output.Cells(k, oX_col + 4).Interior.Color = RGB(217,
217, 217)
    For col = 1 To oX_col + 2
      If output.Cells(k, col).Value = PV_in And PV_in >=
alfa_in Then
        output.Cells(k, col).Interior.Color = RGB(166, 166,
166)
      End If
    Next col
```

```
    With Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).BorderAround (xlDouble)
    Range(output.Cells(1, 2), output.Cells(1, oX_col +
2)).BorderAround (xlDouble)

    Range(output.Cells(1, oX_col + 4), output.Cells(k, oX_col
+ 4)).BorderAround (xlDouble)
    output.Cells(1, oX_col + 4).BorderAround (xlDouble)
    With Range(output.Cells(2, oX_col + 4), output.Cells(k,
oX_col + 4)).Borders(xlInsideHorizontal)
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With

    output.Cells(4, 2) = "ain"
    output.Cells(4, 2).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(4, 2).Characters(Start:=2,
Length:=2).Font.Subscript = True
    output.Cells(4, 2).Font.Bold = True
    output.Cells(4, 3) = alfa_in
    output.Cells(4, 3).NumberFormat = "0.000"

    output.Cells(4, 5) = "aout"
    output.Cells(4, 5).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(4, 5).Characters(Start:=2,
Length:=3).Font.Subscript = True
    output.Cells(4, 5).Font.Bold = True
    output.Cells(4, 6) = alfa_out
    output.Cells(4, 6).NumberFormat = "0.000"

    With Range(output.Cells(4, 2), output.Cells(4,
3)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(4, 2), output.Cells(4,
3)).BorderAround (xlDouble)
    With Range(output.Cells(4, 5), output.Cells(4,
6)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(4, 5), output.Cells(4,
6)).BorderAround (xlDouble)

  Else 'continuamos o algoritmo normalmente ate que a
variavel a entrar deixe de ser significativa mas todas as
restantes sejam
    Do Until PV_in >= alfa_in Or k = oX_col * 3 Or
(X_modelo_col = oX_col And max_PV_out < alfa_in And
max_PV_out < alfa_out)

      For col = 1 To X_col 'UBound(X, 2)
        Call remove_col(X, col, X_temp, col_X)
        Call add_col(X_modelo, col_X, X_teste)

        log_est_output = Log_Est1(y, X_teste, False)

        Rz_Ver =
Application.WorksheetFunction.Index(log_est_output, 1, 1)
        Rz_Ver_output(1, col) = Rz_Ver
      Next col

      min_Rz_Ver =
Application.WorksheetFunction.Min(Rz_Ver_output)
      aux_min_Rz_Ver =
Application.WorksheetFunction.Match(min_Rz_Ver,
Rz_Ver_output, 0)

      'Matriz de planeamento base, com menos uma coluna
      'Matriz de Planeamento com a variavel seleccionada
      If UBound(X, 2) = 1 Then
        For Lin = 1 To oX_lin
          col_X(Lin, 1) = X(Lin, 1)
        Next Lin
        X_col = 0
      Else
        Call remove_col(X, aux_min_Rz_Ver, X, col_X)
        X_col = UBound(X, 2)
      End If
```

```
      Call add_col(X_modelo, col_X, X_modelo)
      X_col = UBound(X, 2)
      X_modelo_col = UBound(X_modelo, 2)

      Call remove_col(H_X, aux_min_Rz_Ver, H_X, col_H)

      'Adicionar a coluna na penultima posicao
      ReDim H_modelo1(1 To 1, 1 To X_modelo_col + 1)
      For col = 1 To X_modelo_col - 1
        H_modelo1(1, col) = H_modelo(1, col)
      Next col
        H_modelo1(1, X_modelo_col) = col_H(1, 1)
        H_modelo1(1, X_modelo_col + 1) = "const"

      ReDim H_modelo(1 To 1, 1 To X_modelo_col + 1)
      For col = 1 To X_modelo_col + 1
        H_modelo(1, col) = H_modelo1(1, col)
      Next col

      'Funcao Log_est (equivalente a LINEST para regressao
logistica)
      log_est_output = Log_Est1(y, X_modelo, True)

      'Para pegar no output do log_est e calcular os P-Values
      Call P_Value_N(X_modelo_col, log_est_output, 1, QF,
P_V, max_P_V, aux_max, PV_in)

      'juntar os p-values no quadro a escrver na folha
      ReDim Q_pv(1 To 1, 1 To oX_col + 1)
      For col = 1 To oX_col + 1
        On Error Resume Next 'para ficar em branco quando o
index-match nao encontrar correspondencia
        Q_pv(1, col) = WorksheetFunction.Index(P_V, 1,
WorksheetFunction.Match(oH(1, col), H_modelo, 0))
      Next col

      'P-Values
      Range(output.Cells(k + 1, 2), output.Cells(k + 1,
oX_col + 2)) = Q_pv
      '-2xlog(Lo/L1)
      output.Cells(k + 1, oX_col + 4) =
Application.WorksheetFunction.Index(log_est_output, 4, 2)
      Range(output.Cells(k + 1, 2), output.Cells(k + 1,
oX_col + 4)).NumberFormat = "#,##0.000"

      For col = 1 To X_modelo_col + 1
        HF(k, col) = H_modelo(1, col)
      Next col

      PV_Col = UBound(P_V, 2) 'tem coluna para o p-value da
constante

      'p-value das restantes
      ReDim PV_out(1 To 1, 1 To PV_Col - 2)
      For col = 1 To PV_Col - 2
        PV_out(1, col) = P_V(1, col)
      Next col
      PV_out_Col = UBound(PV_out, 2)
      max_PV_out = Application.WorksheetFunction.Max(PV_out)

      If max_PV_out >= alfa_out Then 'retiramos as colunas
das variaveis nao significativas
        max_PV_out = 999
        For col = PV_out_Col To 1 Step -1
          If PV_out(1, col) >= alfa_out Then
            Call remove_col(X_modelo, col, X_modelo, col_X)
            Call add_col(X, col_X, X)
            Call remove_col(H_modelo, col, H_modelo, col_H)
            Dim H_X_col_aux As Long
            H_X_col_aux = UBound(H_X, 2)

            'Adicionar a coluna na penultima posicao
            ReDim H_X1(1 To 1, 1 To H_X_col_aux + 1)
            For col2 = 1 To H_X_col_aux - 1
              H_X1(1, col2) = H_X(1, col2)
            Next col2

            H_X1(1, H_X_col_aux) = col_H(1, 1)
            H_X1(1, H_X_col_aux + 1) = "const"

            ReDim H_X(1 To 1, 1 To H_X_col_aux + 1)
            For col2 = 1 To H_X_col_aux + 1
              H_X(1, col2) = H_X1(1, col2)
            Next col2
          Else
          End If
```

```
        Next col
      Else
      End If

      'O modelo final sera o obtido na penultima iteracao
      'guardamos a informacao de cada modelo seleccionado a
cada iteracao
      For col = 1 To X_modelo_col + 1
        QF1(k, col) = QF(1, col)
        QF2(k, col) = QF(2, col)
        QF3(k, col) = QF(3, col)
        QF4(k, col) = QF(4, col)
      Next col

      'Auxiliar para apurar a iteracao com o modelo
seleccionado: sera o ultimo em que todas as variaveis sao
significativas
      ReDim pv_temp(1 To 1, 1 To X_modelo_col - 1)
      For col = 1 To X_modelo_col - 1
        pv_temp(1, col) = QF(4, col)
      Next col
      max_pv_final =
Application.WorksheetFunction.Max(pv_temp)

      If PV_in < alfa_in And max_pv_final < alfa_out Then
        pv_aux_final(k, 1) = 1
      Else
        pv_aux_final(k, 1) = 0
      End If

      Stats1(k, 1) =
Application.WorksheetFunction.Index(log_est_output, 4, 1)
      Stats2(k, 1) =
Application.WorksheetFunction.Index(log_est_output, 4, 2)
      Stats3(k, 1) =
Application.WorksheetFunction.Index(log_est_output, 5, 2)

      'p-value in > alfa_in : variavel testada nao e'
significativa
      For col = 1 To oX_col
        If output.Cells(k, col + 1).Value = "" And
output.Cells(k + 1, col + 1).Value <> "" And output.Cells(k +
1, col + 1).Value >= alfa_in Then
          output.Cells(k + 1, col + 1).Interior.Color =
RGB(166, 166, 166)
        End If
      Next col

      'p-value out > alfa_out : variaveis a sair
      For col = 1 To oX_col
        If output.Cells(k + 1, col + 1).Value > alfa_out Then
          output.Cells(k + 1, col + 1).Interior.Color =
RGB(166, 166, 166)
        End If
      Next col

      'p-value in < alfa_in : variavel testada e'
significativa
      aux = 1
      For col = 1 To oX_col
        If output.Cells(k, col + 1).Value = "" And _
          output.Cells(k + 1, col + 1).Value <> "" And
output.Cells(k + 1, col + 1).Value < alfa_in Then
          output.Cells(k + 1, col + 1).Interior.Color =
RGB(218, 238, 243)
          aux = 0
          col_aux(k + 1, 1) = col
        End If
      Next col

      For col = 1 To oX_col
        If aux = 1 And col <> col_aux(k, 1) And _
          output.Cells(k, col + 1).Value > alfa_out And _
          output.Cells(k + 1, col + 1).Value <> "" And
output.Cells(k + 1, col + 1).Value < alfa_in Then
          output.Cells(k + 1, col + 1).Interior.Color =
RGB(218, 238, 243)
          col_aux(k + 1, 1) = col
        End If
      Next col

      k = k + 1
    Loop

    With Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).Borders
        .LineStyle = xlContinuous
```

```
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(1, 2), output.Cells(k, oX_col +
2)).BorderAround (xlDouble)
    Range(output.Cells(1, 2), output.Cells(1, oX_col +
2)).BorderAround (xlDouble)

    Range(output.Cells(1, oX_col + 4), output.Cells(k, oX_col
+ 4)).BorderAround (xlDouble)
    output.Cells(1, oX_col + 4).BorderAround (xlDouble)
    With Range(output.Cells(2, oX_col + 4), output.Cells(k,
oX_col + 4)).Borders(xlInsideHorizontal)
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With

    'Temos que determinar em que iteracao foi testado o
modelo seleccionado
    iter = 0
    Lin = UBound(pv_aux_final, 1)
    Do Until pv_aux_final(Lin, 1) = 1 Or Lin = 0
      Lin = Lin - 1
      iter = Lin
    Loop

    If iter = 0 Then
      iter = 1
    Else
    End If

    ReDim H_X(1 To 1, 1 To k + 1)
    For col = 1 To k + 1
      H_X(1, col) = HF(iter, col)
    Next col

    X_modelo_col =
Application.WorksheetFunction.Match("const", H_X, 0)

    ReDim QF_final(1 To 4, 1 To X_modelo_col)
    For col = 1 To X_modelo_col + 1
      QF_final(1, col) = QF1(iter, col)
      QF_final(2, col) = QF2(iter, col)
      QF_final(3, col) = QF3(iter, col)
      QF_final(4, col) = QF4(iter, col)
    Next col

    ReDim Stats_final(1 To 3, 1 To 1)
    Stats_final(1, 1) = Stats1(iter, 1)
    Stats_final(2, 1) = Stats2(iter, 1)
    Stats_final(3, 1) = Stats3(iter, 1)

    Range(output.Cells(2, 2), output.Cells(k, oX_col +
4)).NumberFormat = "#,##0.000"


    j = k + 3
    X_col = X_modelo_col

    output.Cells(j + 0, 1) = "Coeficientes"
    output.Cells(j + 1, 1) = "EMQ"
    output.Cells(j + 2, 1) = "s(b)"
    output.Cells(j + 2, 1).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(j + 3, 1) = "Estat. Teste"
    output.Cells(j + 4, 1) = "P-Value"
    Range(output.Cells(j + 0, 2), output.Cells(j + 0, X_col +
1)) = H_X
    Range(output.Cells(j + 0, 1), output.Cells(j + 0, X_col +
1)).Font.Bold = True
    Range(output.Cells(j + 1, 1), output.Cells(j + 4,
1)).Font.Bold = True
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
1)) = QF_final
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
1)).NumberFormat = "#,##0.000"

    With Range(output.Cells(j + 0, 1), output.Cells(j + 4,
X_col + 1)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, 1), output.Cells(j + 4, X_col +
1)).BorderAround (xlDouble)
    output.Cells(j + 0, 1).BorderAround (xlDouble)
    Range(output.Cells(j + 1, 2), output.Cells(j + 4, X_col +
1)).BorderAround (xlDouble)

    output.Cells(j + 0, X_col + 3) = "ain"
```

```
    output.Cells(j + 0, X_col + 3).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(j + 0, X_col + 3).Characters(Start:=2,
Length:=2).Font.Subscript = True
    output.Cells(j + 0, X_col + 3).Font.Bold = True
    output.Cells(j + 0, X_col + 4) = alfa_in
    output.Cells(j + 0, X_col + 4).NumberFormat = "0.000"

    output.Cells(j + 1, X_col + 3) = "aout"
    output.Cells(j + 1, X_col + 3).Characters(Start:=1,
Length:=1).Font.Name = "Symbol"
    output.Cells(j + 1, X_col + 3).Characters(Start:=2,
Length:=3).Font.Subscript = True
    output.Cells(j + 1, X_col + 3).Font.Bold = True
    output.Cells(j + 1, X_col + 4) = alfa_out
    output.Cells(j + 1, X_col + 4).NumberFormat = "0.000"

    With Range(output.Cells(j + 0, X_col + 3), output.Cells(j
+ 1, X_col + 4)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, X_col + 3), output.Cells(j + 1,
X_col + 4)).BorderAround (xlDouble)

    output.Cells(j + 2, X_col + 3) =
Application.WorksheetFunction.Index(log_est_output, 3, 1)
    output.Cells(j + 2, X_col + 3).Characters(Start:=2,
Length:=1).Font.Subscript = True
    output.Cells(j + 2, X_col + 3).Characters(Start:=5,
Length:=1).Font.Subscript = True
    output.Cells(j + 2, X_col + 4) = Stats_final(1, 1)

    output.Cells(j + 3, X_col + 3) =
Application.WorksheetFunction.Index(log_est_output, 3, 2)
    output.Cells(j + 3, X_col + 3).Characters(Start:=8,
Length:=1).Font.Subscript = True
    output.Cells(j + 3, X_col + 3).Characters(Start:=11,
Length:=1).Font.Subscript = True
    output.Cells(j + 3, X_col + 4) = Stats_final(2, 1)

    output.Cells(j + 4, X_col + 3) =
Application.WorksheetFunction.Index(log_est_output, 5, 1)
    output.Cells(j + 4, X_col + 4) = Stats_final(3, 1)
```

```
    With Range(output.Cells(j + 2, X_col + 3), output.Cells(j
+ 4, X_col + 4)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 2, X_col + 3), output.Cells(j + 4,
X_col + 4)).BorderAround (xlDouble)
    Range(output.Cells(j + 2, X_col + 4), output.Cells(j + 3,
X_col + 4)).NumberFormat = "#,##0.000"
    Range(output.Cells(j + 0, X_col + 3), output.Cells(j + 4,
X_col + 3)).Font.Bold = True

    Dim coef As Range
    Set coef = Range(output.Cells(j + 0, 2), output.Cells(j +
1, X_col + 1))

    Call MConfusao_ROC(oY, oX, coef, pc, Y_obs, Y_prev,
ROC_graf)

    output.Cells(j + 0, X_col + 6) = "Cut-off"
    output.Cells(j + 0, X_col + 6).Font.Bold = True
    'por defeito, colocamos o valor medio dos y previstos
    output.Cells(j + 0, X_col + 7) =
WorksheetFunction.Average(Y_prev)
    output.Cells(j + 0, X_col + 7).NumberFormat = "#,##0.000"

    With Range(output.Cells(j + 0, X_col + 6), output.Cells(j
+ 0, X_col + 7)).Borders
    .LineStyle = xlContinuous
    .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 0, X_col + 6), output.Cells(j + 0,
X_col + 7)).BorderAround (xlDouble)
    output.Cells(j + 0, X_col + 7).Interior.Color = RGB(218,
238, 243)

    'Para obter uma matriz de confusao dinamica para qualquer
ponto de corte temos que imprimir
    'os Y Observados e os Y previstos na folha

    Dim aux_col As Long
    aux_col = WorksheetFunction.Max(oX_col + 2, X_col + 11)
```

```
    output.Cells(-2, aux_col + 20).Value = "Y Obs"
    Range(output.Cells(-1, aux_col + 20), output.Cells(oX_lin
- 2, aux_col + 20)) = Y_obs

    output.Cells(-2, aux_col + 21).Value = "Prob"
    Range(output.Cells(-1, aux_col + 21), output.Cells(oX_lin
- 2, aux_col + 21)) = Y_prev

    output.Cells(-2, aux_col + 22).Value = "Cut-off"
    output.Cells(-1, aux_col + 22).FormulaR1C1 = "=R" & k + 6
& "C" & X_col + 8 & ""
    Range(output.Cells(0, aux_col + 22), output.Cells(oX_lin
- 2, aux_col + 22)).FormulaR1C1 = "=R[-1]C[0]"

    output.Cells(-2, aux_col + 23).Value = "Y Prev"
    Range(output.Cells(-1, aux_col + 23), output.Cells(oX_lin
- 2, aux_col + 23)).FormulaR1C1 = "=IF(R[0]C[-2]>R[0]C[-
1],1,0)"

    output.Cells(-2, aux_col + 24).Value = "O1 + P1"
    Range(output.Cells(-1, aux_col + 24), output.Cells(oX_lin
- 2, aux_col + 24)).FormulaR1C1 = "=if(and(R[0]C[-
4]=1,R[0]C[-1]=1),1,0)"

    output.Cells(-2, aux_col + 25).Value = "O0 + P1"
    Range(output.Cells(-1, aux_col + 25), output.Cells(oX_lin
- 2, aux_col + 25)).FormulaR1C1 = "=if(and(R[0]C[-
5]=0,R[0]C[-2]=1),1,0)"

    output.Cells(-2, aux_col + 26).Value = "O1 + P0"
    Range(output.Cells(-1, aux_col + 26), output.Cells(oX_lin
- 2, aux_col + 26)).FormulaR1C1 = "=if(and(R[0]C[-
6]=1,R[0]C[-3]=0),1,0)"

    output.Cells(-2, aux_col + 27).Value = "O0 + P0"
    Range(output.Cells(-1, aux_col + 27), output.Cells(oX_lin
- 2, aux_col + 27)).FormulaR1C1 = "=if(and(R[0]C[-
7]=0,R[0]C[-4]=0),1,0)"

    output.Cells(j + 2, X_col + 8).Value = "Observados"
    Range(output.Cells(j + 2, X_col + 8), output.Cells(j + 2,
X_col + 9)).Merge
    output.Cells(j + 3, X_col + 8).Value = "Positivos"
    output.Cells(j + 3, X_col + 9).Value = "Negativos"
    output.Cells(j + 3, X_col + 10).Value = "Total"

    output.Cells(j + 4, X_col + 6).Value = "Previstos"
    Range(output.Cells(j + 4, X_col + 6), output.Cells(j + 5,
X_col + 6)).Merge
    output.Cells(j + 4, X_col + 7).Value = "Positivos"
    output.Cells(j + 5, X_col + 7).Value = "Negativos"
    output.Cells(j + 6, X_col + 7).Value = "Total"

    'Verdadeiros Positivos: Observado=1 e Previsto=1
    output.Cells(j + 4, X_col + 8).FormulaR1C1 = "=sum(R" & 2
& "C" & aux_col + 25 & ":R" & oX_lin + 1 & "C" & aux_col + 25
& ")"

    'Falsos Negativos: Observado=1 e Previsto=0
    output.Cells(j + 4, X_col + 9).FormulaR1C1 = "=sum(R" & 2
& "C" & aux_col + 26 & ":R" & oX_lin + 1 & "C" & aux_col + 26
& ")"

    'Falsos Positivos: Observado=0 e Previsto=1
    output.Cells(j + 5, X_col + 8).FormulaR1C1 = "=sum(R" & 2
& "C" & aux_col + 27 & ":R" & oX_lin + 1 & "C" & aux_col + 27
& ")"

    'Verdadeiros Negativos= Observado=0 e Previsto=0
    output.Cells(j + 5, X_col + 9).FormulaR1C1 = "=sum(R" & 2
& "C" & aux_col + 28 & ":R" & oX_lin + 1 & "C" & aux_col + 28
& ")"

    'Totais
    Range(output.Cells(j + 6, X_col + 8), output.Cells(j + 6,
X_col + 9)) = "=sum(R[" & -2 & "]C[0]:R[" & -1 & "]C[0])"
    Range(output.Cells(j + 4, X_col + 10), output.Cells(j +
5, X_col + 10)) = "=sum(R[0]C[" & -2 & "]:R[0]C[" & -1 & "])"
    output.Cells(j + 6, X_col + 10) = "=sum(R[" & -2 & "]C["
& -2 & "]:R[" & -1 & "]C[" & -1 & "])"


    Range(output.Cells(j + 2, X_col + 8), output.Cells(j + 2,
X_col + 9)).BorderAround (xlDouble)
    Range(output.Cells(j + 2, X_col + 8), output.Cells(j + 2,
X_col + 10)).Font.Bold = True
```

```
    With Range(output.Cells(j + 3, X_col + 8), output.Cells(j
+ 3, X_col + 9)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 3, X_col + 8), output.Cells(j + 3,
X_col + 9)).BorderAround (xlDouble)
    Range(output.Cells(j + 3, X_col + 8), output.Cells(j + 3,
X_col + 9)).Font.Bold = True

    Range(output.Cells(j + 4, X_col + 6), output.Cells(j + 5,
X_col + 6)).BorderAround (xlDouble)
    Range(output.Cells(j + 4, X_col + 6), output.Cells(j + 5,
X_col + 6)).Font.Bold = True

    With Range(output.Cells(j + 4, X_col + 7), output.Cells(j
+ 5, X_col + 7)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 4, X_col + 7), output.Cells(j + 5,
X_col + 7)).BorderAround (xlDouble)
    Range(output.Cells(j + 4, X_col + 7), output.Cells(j + 6,
X_col + 7)).Font.Bold = True

    With Range(output.Cells(j + 4, X_col + 8), output.Cells(j
+ 5, X_col + 9)).Borders
        .LineStyle = xlContinuous
        .Color = RGB(166, 166, 166)
    End With
    Range(output.Cells(j + 4, X_col + 8), output.Cells(j + 5,
X_col + 9)).BorderAround (xlDouble)

    Range(output.Cells(j + 6, X_col + 7), output.Cells(j + 6,
X_col + 10)).Font.Color = RGB(89, 89, 89)
    Range(output.Cells(j + 3, X_col + 10), output.Cells(j +
5, X_col + 10)).Font.Color = RGB(89, 89, 89)
    output.Cells(j + 6, X_col + 10).Font.Bold = True

    With output.Cells(j + 6, X_col + 10).Borders(xlEdgeTop)
        .LineStyle = xlDouble
        .Color = RGB(166, 166, 166)
    End With
    With output.Cells(j + 6, X_col + 10).Borders(xlEdgeLeft)
        .LineStyle = xlDouble
        .Color = RGB(166, 166, 166)
    End With

    'Print curva ROC
    '1a coluna: pontos de corte
    '2a coluna: 1-E = 1 - (VN/(VN+FP))
    '3a coluna: S = VP/(VP+FN)

    output.Cells(j + 0, X_col + 14).Value = "Dados Curva ROC"
    Range(output.Cells(j + 0, X_col + 14), output.Cells(j +
0, X_col + 16)).Merge
    output.Cells(j + 1, X_col + 14).Value = "pt corte"
    output.Cells(j + 1, X_col + 15).Value = "1-E"
    output.Cells(j + 1, X_col + 16).Value = "S"
    Range(output.Cells(j + 0, X_col + 14), output.Cells(j +
1, X_col + 16)).Font.Bold = True

    Range(output.Cells(j + 2, X_col + 14), output.Cells(j +
pc + 1, X_col + 16)) = ROC_graf
    Range(output.Cells(j + 2, X_col + 14), output.Cells(j +
pc + 1, X_col + 16)).NumberFormat = "0.00"

    Call graph_ROC(output, j, X_col, pc)
    aux_ROC = 1

  End If

  output.Cells(j + 6, 1).Value = "Seleção Y:"
  output.Cells(j + 6, 1).Font.Bold = True
  output.Cells(j + 7, 1).Value = oY.Worksheet.Name
  output.Cells(j + 8, 1).Value = oY.Address(False, False)

  output.Cells(j + 10, 1).Value = "Seleção X:"
  output.Cells(j + 10, 1).Font.Bold = True
  output.Cells(j + 11, 1).Value = oX.Worksheet.Name
  output.Cells(j + 12, 1).Value = oX.Address(False, False)

  Sheets("Out_Stepwise_Logistica").Cells.EntireColumn.AutoFit
  Sheets("Out_Stepwise_Logistica").Cells.EntireRow.AutoFit

  If aux_ROC = 1 Then
    output.Cells(-2, oX_col + 6).Value = "ATENÇÃO: recomenda-
se apagar o gráfico antes de apagar a folha!"
```

```
    output.Cells(-2, oX_col + 6).HorizontalAlignment = xlLeft
    output.Cells(-2, oX_col + 6).Characters(Start:=1,
Length:=8).Font.Bold = True
    output.Cells(-2, oX_col + 6).Characters(Start:=1,
Length:=8).Font.Color = RGB(255, 0, 0)
  End If

End Sub
```

11.10. SUB_10_CORRELACOES

```
'Inputs:
'  - oX: matriz original (matriz de planeamento);
'Outputs:
'  - MCorrelacao_inv: matriz inversa da matriz de correlações
da matriz oX;
'  - max_diagonal: maior valor da diagonal principal da
matriz MCorrelacao_inv;
'  - ordem_max: número da coluna correspondente ao elemento
de maior valor na diagonal da matriz MCorrelacao_inv.
'-------------------------------------------------------------
Sub correlacoes(oX, MCorrelacao_inv, max_diagonal, ordem_max)

  Dim oX_col As Long, oX_lin As Long
  X_lin = UBound(oX, 1)
  X_col = UBound(oX, 2)

  ReDim X(1 To X_lin, 1 To X_col)
  For col = 1 To X_col
    For Lin = 1 To X_lin
      X(Lin, col) = oX(Lin, col)
    Next Lin
  Next col

  ReDim MCorrelacao(1 To X_col, 1 To X_col)
    For i = 1 To X_col
    MCorrelacao(i, i) = 1
    For j = i + 1 To X_col

      ReDim X_coluna_i(1 To X_lin, 1 To 1), X_coluna_j(1 To
X_lin, 1 To 1)
      For Lin = 1 To X_lin
        X_coluna_i(Lin, 1) = X(Lin, i)
        X_coluna_j(Lin, 1) = X(Lin, j)
      Next Lin

      MCorrelacao(i, j) =
WorksheetFunction.Correl(X_coluna_i, X_coluna_j)
      MCorrelacao(j, i) = MCorrelacao(i, j)
    Next j
  Next i

  'Temos que arredondar os valores da matriz de correlacoes a
10 casas decimais, pois em situacoes extremas,
  'com mais casas decimais, o excel nao inverte a matriz
corretamente
  ReDim MCorrelacao_arrend(1 To X_col, 1 To X_col)
  For Lin = 1 To X_col
    For col = 1 To X_col
      MCorrelacao_arrend(Lin, col) =
WorksheetFunction.Round(MCorrelacao(Lin, col), 10)
    Next col
  Next Lin

  ReDim MCorrelacao_inv(1 To X_col, 1 To X_col)
  MCorrelacao_inv =
WorksheetFunction.MInverse(MCorrelacao_arrend)

  ReDim diagonal(1 To 1, 1 To X_col)
  For col = 1 To X_col
    diagonal(1, col) = MCorrelacao_inv(col, col)
  Next col

  max_diagonal = WorksheetFunction.Max(diagonal)
  ordem_max =
Application.WorksheetFunction.Match(max_diagonal, diagonal,
0)

End Sub
```

11.11. SUB_11_ADD_COL

```
'Inputs:
'  - oM: matriz original;
'  - M_col: coluna a adicionar no fim da matriz original;
'Outputs:
'  - M: nova matriz com mais uma coluna que oM.
'-------------------------------------------------------
```

```
Sub add_col(oM, M_col, M)

  Dim rw As Long, cl As Long, result()
  rw = UBound(oM, 1)
  cl = UBound(oM, 2)

  ReDim result(1 To rw, 1 To cl + 1)
  For Lin = 1 To rw
    For col = 1 To cl
      result(Lin, col) = oM(Lin, col)
    Next col
      result(Lin, cl + 1) = M_col(Lin, 1)
  Next Lin
  M = result()
End Sub
```

11.12. SUB_12_REMOVE_COL

```
'Inputs:
'   - oM: matriz original;
'   - k: número da coluna a retirar;
'Outputs:
'   - M: nova matriz com menos uma coluna que oM;
'   - col_M: coluna de oM que foi retirada.
'-------------------------------------------------------------
--
Sub remove_col(oM, k, M, col_M)

  Dim rw As Long, cl As Long, result()
  rw = UBound(oM, 1)
  cl = UBound(oM, 2)

  If k = 0 Then
    ReDim result(1 To rw, 1 To cl), col_M(1 To rw, 1 To 1)
    For Lin = 1 To rw
      For col = 1 To cl
        result(Lin, col) = oM(Lin, col)
      Next col
        col_M(Lin, 1) = 0
    Next Lin

  Else
    ReDim result(1 To rw, 1 To cl - 1), col_M(1 To rw, 1 To
1)
    For Lin = 1 To rw
      col_M(Lin, 1) = oM(Lin, k)
    Next Lin

    col_aux = 1
    For col = 1 To cl - 1

      If col = k Then
        col_aux = col_aux + 1
      Else
        col_aux = col_aux
      End If

      For Lin = 1 To rw
        result(Lin, col) = oM(Lin, col_aux)
      Next Lin

    col_aux = col_aux + 1
    Next col
  End If
  M = result()
End Sub
```

11.13. SUB_13_LOG_EST

```
Function Log_Est1(y, oX, Out As Boolean)

  Dim N As Long, p As Long, Dist As Double, Media As Double,
Ajust As Double
  Dim B(), eta(), U(), X(), result()
  Dim XTWX(), XTWZ(), XTWXInv()

  'Obter a dimensao da amostra e o numero de variaveis
independentes
  N = UBound(oX, 1)
  p = UBound(oX, 2)
  p = p + 1

  ReDim X(1 To N, 1 To p), B(1 To p, 1 To 1)
  ReDim eta(1 To N), U(1 To p), result(1 To 5, 1 To p)
  ReDim XTWX(1 To p, 1 To p), XTWZ(1 To p, 1 To 1), XTWXInv(1
To p, 1 To p)
```

```
  'Copiar as celas na seleccao para o array X e Y e calcular
a Media
  Media = 0
  For i = 1 To N
    Media = Media + y(i, 1)
  Next i
    Media = Media / N

  For i = 1 To N
    X(i, 1) = 1
  Next i
  For i = 1 To N
    For j = 2 To p
      X(i, j) = oX(i, j - 1)
    Next j
  Next i

  'Iniciar o vector de parametros B
  B(1, 1) = Log(Media / (1 - Media))
  For i = 2 To p
  B(i, 1) = 0
  Next i

  Do
  For i = 1 To N
    eta(i) = 0
    For j = 1 To p
      eta(i) = eta(i) + B(j, 1) * X(i, j)
    Next j
  Next i


  'Construir as matrizes XTWX, XTWXInv e XTWZ
  For j = 1 To p
    For k = 1 To p
      XTWX(j, k) = 0
      For i = 1 To N
        XTWX(j, k) = XTWX(j, k) + X(i, j) * Exp(eta(i)) *
X(i, k) / (1 + Exp(eta(i))) ^ 2
      Next i
    Next k
  Next j

  XTWXInv = Application.MInverse(XTWX)

  For j = 1 To p
    XTWZ(j, 1) = 0
    For i = 1 To N
      XTWZ(j, 1) = XTWZ(j, 1) + X(i, j) * (Exp(-eta(i)) *
eta(i) / (1 + Exp(-eta(i))) ^ 2 + (y(i, 1) - 1 / (1 + Exp(-
eta(i)))))
    Next i
  Next j

  'Actualizar estimador dos parametros B
  B = Application.MMult(XTWXInv, XTWZ)

  'Calcular U(B)
  For j = 1 To p
    Dist = 0
    U(j) = 0
    For i = 1 To N
      U(j) = U(j) + (y(i, 1) - 1 / (1 + Exp(-eta(i)))) * X(i,
j)
    Next i
    Dist = Dist + U(j) ^ 2
  Next j

  Loop Until Dist < 0.0001

  'Calcular Ajustamento
  Ajust = 0
  For i = 1 To N
    Ajust = Ajust + y(i, 1) * Log(Media * (1 + Exp(-eta(i))))
+ (1 - y(i, 1)) * Log((1 - Media) * (1 + Exp(eta(i))))
  Next i
    Ajust = Exp(Ajust)

  'Colocamos os estimadores da contante no fim do vetor para
ficar alinhado com o output da funcao LINEST
  ReDim B_aux(1 To p, 1 To 1), XTWXInv_aux(1 To p, 1 To 1)

  'pomos os estimadores da constante no fim
  For Lin = 1 To p - 1
    B_aux(Lin, 1) = B(Lin + 1, 1)
  Next Lin
    B_aux(p, 1) = B(1, 1)
```

```
  For Lin = 1 To p - 1
    XTWXInv_aux(Lin, 1) = (XTWXInv(Lin + 1, Lin + 1)) ^ (1 /
2)
  Next Lin
    XTWXInv_aux(p, 1) = (XTWXInv(1, 1)) ^ (1 / 2)

  If Out = True Then
    For j = 1 To p
    result(1, j) = B_aux(j, 1)
    result(2, j) = XTWXInv_aux(j, 1)
    Next j

    result(3, 1) = "L0/L1"
    result(3, 2) = "'-2*ln(L0/L1)"
    result(4, 1) = Ajust
    result(4, 2) = -2 * Log(Ajust)

    For j = 3 To p
      result(3, j) = "#N/A"
      result(4, j) = "#N/A"
    Next j
    result(5, 1) = "G.L."
    result(5, 2) = p - 1

    For j = 3 To p
      result(5, j) = "#N/A"
    Next j
  Else
    result(1, 1) = Ajust
    result(2, 1) = -2 * Log(Ajust)
  End If

  Log_Est1 = result()

End Function
```

11.14. SUB_14_P_VALUE_T

```
'Input:
'  - X_col: número de variáveis independentes da iteração;
'  - linest_output: output da função LINEST;
'  - ct: indicador de existência de constante no modelo
final;
'Output:
'  - QF: Quadro final com a infomação dos cálculos da
iteração;
'  - P_V: vetor de P-Values (nota: não podemos usar "PV" como
variável pois é uma função do VBA);
'  - max_P_V: maior p-value da iteração;
'  - aux_max: número da coluna correspondente do maior p-
value;
'  - P_V_ult: p-value da última variável independente.
'-------------------------------------------------------------

Sub P_Value_T(X_col, linest_output, ct, QF, P_V, max_P_V,
aux_max, P_V_ult)
  Dim GL As Long
  ReDim QF(1 To 4, 1 To X_col + 1), P_V(1 To 1, 1 To X_col +
1)

  'Para aceder as posicoes do output da funcao LINEST temos
que usar a funcao INDEX
  '1a linha: Estimadores de Minimos Quadrados
  '2a linha: Desvio-padrao

  For Lin = 1 To 2
    For col = 1 To X_col
      On Error Resume Next 'para ficar em branco quando o
index
      QF(Lin, col) =
Application.WorksheetFunction.Index(linest_output, Lin, X_col
- col + 1)
    Next col
      On Error Resume Next 'para ficar em branco quando o
index
      QF(Lin, X_col + 1) =
Application.WorksheetFunction.Index(linest_output, Lin, X_col
+ 1)
  Next Lin


  'GL: Graus de Liberdade -> celula [4,2] do output da funcao
LINEST
  GL = Application.WorksheetFunction.Index(linest_output, 4,
2)

  'ET_T: Estatistica de Teste do Teste-T -> EMQ / Des.P
```

```
  For col = 1 To X_col + ct
    QF(3, col) = QF(1, col) / QF(2, col)
  Next col


  'P_V: P-Value - TDist(<Est. Teste>, GL, <n° caudas>)
  For col = 1 To X_col + ct
    QF(4, col) =
Application.WorksheetFunction.TDist(Abs(QF(3, col)), GL, 2)
    P_V(1, col) =
Application.WorksheetFunction.TDist(Abs(QF(3, col)), GL, 2)
  Next col

  'variavel com a menor significancia = Maior p-value
  'Guardamos o numero da coluna da variavavel com menor
significancia
  'NOTA: ignoramos o p-value da constante (que estara no fim)

   ReDim P_V_aux(1 To 1, 1 To X_col)
   For col = 1 To X_col
     P_V_aux(1, col) = P_V(1, col)
   Next col

   max_P_V = Application.WorksheetFunction.Max(P_V_aux)
   aux_max = Application.WorksheetFunction.Match(max_P_V,
P_V_aux, 0)

   P_V_ult = P_V(1, X_col)

End Sub
```

11.15. SUB_15_P_VALUE_N

```
'Input:
'  - X_col: número de variáveis independentes da iteração;
'  - log_est_output: output da função LOG_EST1;
'  - ct: indicador de existência de constante no modelo
final;
'Output:
'  - QF: Quadro final com a infomação dos cálculos da
iteração;
'  - P_V: vetor de P-Values (nota: não podemos usar "PV" como
variável pois é uma função do VBA);
'  - max_P_V: maior p-value da iteração;
'  - aux_max: número da coluna correspondente do maior p-
value;
'  - P_V_ult: p-value da última variável independente.
'-------------------------------------------------------------
Sub P_Value_N(X_col, log_est_output, ct, QF, P_V, max_P_V,
aux_max, P_V_ult)
  Dim GL As Long
  ReDim QF(1 To 4, 1 To X_col + 1), P_V(1 To 1, 1 To X_col +
1)

  'Para aceder as posicoes do output da funcao LOG_EST temos
que usar a funcao INDEX
  'As variaveis ja saem na ordem certa, ao contrario da
funcao LINEST
  '1a linha: Estimadores de Minimos Quadrados
  '2a linha: Desvio-padrao

  For Lin = 1 To 2
    For col = 1 To X_col + 1
      On Error Resume Next 'para ficar em branco quando o
index
      QF(Lin, col) =
Application.WorksheetFunction.Index(log_est_output, Lin, col)
    Next col
  Next Lin

  'ET_T: Estatistica de Teste do Teste-N -> EMV / Des.P
  For col = 1 To X_col + ct
    QF(3, col) = QF(1, col) / QF(2, col)
  Next col


  'P_V: P-Value - TDist(<Est. Teste>, GL, <n° caudas>)
  For col = 1 To X_col + ct
    QF(4, col) = 2 * (1 -
Application.WorksheetFunction.Norm_S_Dist(Abs(QF(3, col)),
True))
    P_V(1, col) = 2 * (1 -
Application.WorksheetFunction.Norm_S_Dist(Abs(QF(3, col)),
True))

  Next col

  'variavel com a menor significancia = Maior p-value
```

```
  'Guardamos o numero da coluna da variavavel com menor
significancia
  'NOTA: ignoramos o p-value da constante (que estara no fim)

  ReDim P_V_aux(1 To 1, 1 To X_col)
  For col = 1 To X_col
    P_V_aux(1, col) = P_V(1, col)
  Next col

  max_P_V = Application.WorksheetFunction.Max(P_V_aux)
  aux_max = Application.WorksheetFunction.Match(max_P_V,
P_V_aux, 0)

  P_V_ult = P_V(1, X_col)

End Sub
```

11.16. SUB_16_MCONFUSAO_ROC

```
'Input:
'  - oY: Variável Dependente;
'  - oX: Variáveis Independenes do modelo final;
'  - coef: Vetor dos coeficientes do modelo final;
'  - N: Número de pontos a representar na curva ROC;
'Output:
'  - y: Variável Dependente;
'  - Y_prev: valores previstos da variável Dependente;
'  - ROC_graph: matriz auxiliar com os valores necessários
para desenhar a curva ROC
'-------------------------------------------------------------
Sub MConfusao_ROC(oY, oX, coef, N As Long, y, Y_prev,
ROC_graf)

  'Vamos calcular os valores previstos
  X_lin = oX.Rows.Count - 1
  X_col = oX.Columns.Count
  coef_col = coef.Columns.Count

  ReDim y(1 To X_lin, 1 To 1), X(1 To X_lin, 1 To X_col)

  For Lin = 1 To X_lin
    y(Lin, 1) = oY(Lin + 1, 1)
  Next Lin

  For col = 1 To X_col
    For Lin = 1 To X_lin
      X(Lin, col) = oX(Lin + 1, col)
    Next Lin
  Next col

  ReDim X_coef(1 To X_lin, 1 To X_col + 1)
  For Lin = 1 To X_lin
    For col = 1 To X_col
      If Not IsError(Application.HLookup(oX(1, col), coef, 2,
0)) Then
        X_coef(Lin, col) = X(Lin, col) *
WorksheetFunction.HLookup(oX(1, col), coef, 2, 0)
      Else
        X_coef(Lin, col) = X(Lin, col) * 0
      End If
    Next col
    X_coef(Lin, X_col + 1) = coef(2, coef_col) 'Coeficiente
da Constante
  Next Lin

  ReDim Y_previsto_aux(1 To X_lin, 1 To 1)
  For Lin = 1 To X_lin
    For col = 1 To X_col + 1
      Y_previsto_aux(Lin, 1) = Y_previsto_aux(Lin, 1) +
X_coef(Lin, col) 'Para somar todas as variáveis devidamente
ponderadas pelos seus coeficientes
    Next col
  Next Lin

  ReDim Y_prev(1 To X_lin, 1 To 1)
  For Lin = 1 To X_lin
    Y_prev(Lin, 1) = Exp(Y_previsto_aux(Lin, 1)) / (1 +
Exp(Y_previsto_aux(Lin, 1)))
  Next Lin

  'Pontos de corte
  'Fazemos sempre com um número fixo de pontos com intervalos
iguais entre o mínimo e o máximo dos valores previstos
  Dim min_Y_prev As Double, max_Y_prev As Double, h As Double
  min_Y_prev = Application.WorksheetFunction.Min(Y_prev)
  max_Y_prev = Application.WorksheetFunction.Max(Y_prev)

  h = (max_Y_prev - min_Y_prev) / (N - 1)
```

```
  ReDim pts_corte(1 To N, 1 To 1)
  pts_corte(1, 1) = max_Y_prev
  For Lin = 2 To N - 1
    pts_corte(Lin, 1) = pts_corte(Lin - 1, 1) - h
  Next Lin
    pts_corte(N, 1) = pts_corte(N - 1, 1) - h - 0.000000001

  'Usando os pontos de corte vamos fazer ciclos para obter a
matriz de confusao para cada pt de corte
  ReDim Y_ajust(1 To X_lin, 1 To N)
  For col = 1 To N
    For Lin = 1 To X_lin
      If Y_prev(Lin, 1) > pts_corte(col, 1) Then
        Y_ajust(Lin, col) = 1
      Else
        Y_ajust(Lin, col) = 0
      End If
    Next Lin
  Next col

  ReDim ROC_aux(1 To N, 1 To 4)
  For pc = 1 To N
    ReDim Y_ajust_aux(1 To X_lin, 1 To 1)
    For Lin = 1 To X_lin
      Y_ajust_aux(Lin, 1) = Y_ajust(Lin, pc)
    Next Lin

    'Verdadeiros Positivos: Observado=1 e Previsto=1
    VP = 0
    For Lin = 1 To X_lin
      If y(Lin, 1) = 1 And Y_ajust_aux(Lin, 1) = 1 Then
        VP = VP + 1
      Else
        VP = VP
      End If
    Next Lin
    ROC_aux(pc, 1) = VP

    'Falsos Negativos: Observado=1 e Previsto=0
    FN = 0
    For Lin = 1 To X_lin
      If y(Lin, 1) = 1 And Y_ajust_aux(Lin, 1) = 0 Then
        FN = FN + 1
      Else
        FN = FN
      End If
    Next Lin
    ROC_aux(pc, 2) = FN


    'Falsos Positivos: Observado=0 e Previsto=1
    FP = 0
    For Lin = 1 To X_lin
      If y(Lin, 1) = 0 And Y_ajust_aux(Lin, 1) = 1 Then
        FP = FP + 1
      Else
        FP = FP
      End If
    Next Lin
    ROC_aux(pc, 3) = FP

    'Verdadeiros Negativos= Observado=0 e Previsto=0
    VN = 0
    For Lin = 1 To X_lin
      If y(Lin, 1) = 0 And Y_ajust_aux(Lin, 1) = 0 Then
        VN = VN + 1
      Else
        VN = VN
      End If
    Next Lin
    ROC_aux(pc, 4) = VN
  Next pc

  'Tabela para desenhar a curva ROC
  '1a coluna: Pontos de Corte
  '2a coluna: 1-E = 1 - (VN/(VN+FP))
  '3a coluna: S = VP/(VP+FN)
  ReDim ROC_graf(1 To N, 1 To 3)

  For Lin = 1 To N
    ROC_graf(Lin, 1) = pts_corte(Lin, 1)
    ROC_graf(Lin, 2) = 1 - (ROC_aux(Lin, 4) / (ROC_aux(Lin,
4) + ROC_aux(Lin, 3)))
    ROC_graf(Lin, 3) = ROC_aux(Lin, 1) / (ROC_aux(Lin, 1) +
ROC_aux(Lin, 2))
  Next Lin
End Sub
```

11.17. SUB_17_GRAPH_ROC

```
Sub graph_ROC(output, j, X_col, pc)

  Dim chrt As Chart
  Set chrt = Sheets(output.Worksheet.Name).Shapes.AddChart.Chart

    With chrt
        .ChartType = xlXYScatterLinesNoMarkers

        .SeriesCollection.NewSeries
        .SeriesCollection(1).Name = "=""S"""
        .SeriesCollection(1).XValues = Range(output.Cells(j + 2,
X_col + 15), output.Cells(j + pc + 1, X_col + 15))
        .SeriesCollection(1).Values = Range(output.Cells(j + 2, X_col
+ 16), output.Cells(j + pc + 1, X col + 16))


        .SeriesCollection(2).Name = "=""1-E"""
        .SeriesCollection(2).XValues = Range(output.Cells(j + 2,
X_col + 15), output.Cells(j + pc + 1, X_col + 15))
        .SeriesCollection(2).Values = Range(output.Cells(j + 2, X col
+ 15), output.Cells(j + pc + 1, X col + 15))

        .Axes(xlCategory).Select
        .Axes(xlCategory).MinimumScale = 0
        .Axes(xlCategory).MaximumScale = 1
        .Axes(xlCategory).MajorUnit = 0.1

        .Axes(xlValue).Select
        .Axes(xlValue).MinimumScale = 0
        .Axes(xlValue).MaximumScale = 1
        .Axes(xlValue).MajorUnit = 0.1

        If Application.International(xlDecimalSeparator) = "." Then
          .Axes(xlCategory).TickLabels.NumberFormat = "0.0"
          .Axes(xlValue).TickLabels.NumberFormat = "0.0"
        ElseIf Application.International(xlDecimalSeparator) = ","
Then
          .Axes(xlCategory).TickLabels.NumberFormat = "0,0"
          .Axes(xlValue).TickLabels.NumberFormat = "0,0"
        End If

        .Legend.Delete
        With chrt.Axes(xlCategory)
         .HasTitle = True
         .AxisTitle.Text = "Especificidade (1-E)"
        End With
        With chrt.Axes(xlValue)
         .HasTitle = True
         .AxisTitle.Text = "Sensibilidade (S)"
        End With

       .ChartArea.Select
        With Selection.Format.Line
          .Visible = msoTrue
          .ForeColor.RGB = RGB(166, 166, 166)
        End With

        .Axes(xlValue).HasMajorGridlines = True
        .Axes(xlValue).MajorGridlines.Select
        With Selection.Format.Line
          .Visible = msoTrue
          .ForeColor.RGB = RGB(166, 166, 166)
        End With

        .Axes(xlCategory).HasMajorGridlines = True
        .Axes(xlCategory).MajorGridlines.Select
        With Selection.Format.Line
          .Visible = msoTrue
          .ForeColor.RGB = RGB(166, 166, 166)
        End With

        .PlotArea.Select
        With Selection.Format.Line
          .Visible = msoTrue
          .ForeColor.RGB = RGB(166, 166, 166)
        End With

        .SeriesCollection(1).Select 'FullSeriesCollection (2013)
        With Selection.Format.Line
          .ForeColor.RGB = RGB(44, 133, 156)
          .Visible = msoTrue
          .Weight = 3
        End With

       .SeriesCollection(2).Select
       With Selection.Format.Line
         .Visible = msoTrue
         .DashStyle = msoLineSysDot
         .ForeColor.RGB = RGB(0, 0, 0)
         .Weight = 3
       End With

    End With

End Sub
```